Influence Measures for Logistic Regression, Another Point of View

by

Wesley O. Johnson
University of California at Davis and
University of Minnesota
Technical Report No. 420

Research supported by NIH grant GM25271.

ABSTRACT

Several measures of influence for logistic regression have been suggested. These measures have been developed for the purpose of identifying observations which are influential relative to the estimation of the regression coefficients vector and the deviance. We propose measures for detecting influence relative to the determination of probabilities and the classification of future observations. The relationships among measures are indicated.

Key Words: Logistic regression, influence, prediction, classification.

1.0 Introduction

In this paper we derive and study statistics for detecting and characterizing influential observations in logistic regression. Several of the statistics we consider have been discussed in great detail by Pregibon (1981). While his focus is on measuring the effects observations have on the estimation of the regression coefficients vector and on a particular goodness of fit measure, we focus on the effects observations have on the determination of probabilities and on the classification of future observations. This approach is reasonable since ultimately, the fitted logistic regression (LR) model must be employed with these goals in mind. In the process of developing new methods, and in considering existing methods from this alternative point of view, we are able to provide some new insight. In particular, we show that the Cook (1977) adaptation for detecting influence (discussed in Pregibon (1981) and Cook and Weisberg (1982)) may be interpreted as an approximation to measures of influence relative to Pregibon's goals and ours as well.

In most examples considered we find that the Cook adaptation is a "good" measure of detecting influence and that it is useful for interpretive purposes to a point. However, any instrument employed simultaneously for many different purposes is necessarily "blunt". The additional measures suggested by Pregibon (1981) and the measures we provide, are thus useful for "fine tuning". For aesthetic reasons, we of course prefer our approach to influence, however since some of our measures can be more expensive to compute, we will recommend a middle road whereby observations are initially detected by selected "inexpensive" influence measures, after which other measures are calculated for only the most influential cases.

Cook and Weisberg (1982) provide a general review of work on existing methods of detecting influential observations, including a review of the methods of Pregibon (1981), which are adaptable to the generalized linear model c.f. Nelder and Wedderburn (1972). The work of Johnson and Geisser (1981, 1982, 1983) focuses on the detection of influencial cases relative to the goals of prediction and estimation in the normal theory linear model. Their methods are adaptable to much broader statistical paradigms and are adapted to some extent in this paper.

In section 2 we define the LR model, discuss appropriate inferential goals and corresponding notions of influence after which we discuss some technical results necessary later on. In section 3 we focus on the definitions and interpretations of the various influence measures. In section 4 we consider examples and conclusions are provided in section 5.

2.0 Preliminaries

2.1 General Setting

We begin by making basic definitions and by discussing results which will be needed in later sections. Related discussions may be found in Pregibon (1981) and Cook and Weisberg (1982).

We assume a sample of observations on N individuals $\{(y_1,\underset{\sim}{x}_1),\ldots,(y_N,\underset{\sim}{x}_N)\}$ which have been independently observed. The $\underset{\sim}{x}_i$'s are $1\times p$ vectors of covariates (the first coordinate being a one when a constant is included) and the y_i's are assumed to be realizations of Binomial (n_i,p_i) random variables where

$$p_i = p(\underset{\sim}{x}_i\underset{\sim}{\beta}) = \exp(\underset{\sim}{x}_i\underset{\sim}{\beta})/(1+\exp(\underset{\sim}{x}_i\underset{\sim}{\beta}))$$

and $\underset{\sim}{\beta}$ is a $p\times 1$ vector of unknown regression coefficients. This defines the logistic regression model. For future reference, we will denote the population corresponding to "successes" as π_1 and that corresponding to "failures" as π_0.

The log likelihood function for $\underset{\sim}{\beta}$ may be expressed as

$$\ell(X\underset{\sim}{\beta}) = \sum_{j=1}^{N} \{y_j \underset{\sim}{x}_j \underset{\sim}{\beta} - \ln(1+\exp(\underset{\sim}{x}_j \underset{\sim}{\beta}))\}.$$

The maximum likelihood estimate, $\hat{\underset{\sim}{\beta}}$, may be obtained by solving the likelihood equations

$$X'(\underset{\sim}{y} - \hat{\underset{\sim}{y}}) = 0$$

Where $\underset{\sim}{y}' = (y_1,\ldots,y_N)$, $X' = (\underset{\sim}{x}_1',\ldots,\underset{\sim}{x}_N')$, $\hat{\underset{\sim}{y}}' = (\hat{y}_1,\ldots,\hat{y}_N)$ and $\hat{y}_j = n_j\hat{p}_j = n_j\, p(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}})$ for $j = 1,\ldots,N$. Define the weights $\hat{w}_j = n_j\hat{p}_j(1-\hat{p}_j)$, and the diagonal matrix $\hat{W} = \mathrm{diag}\{\hat{w}_1,\ldots,\hat{w}_N\}$. Then standard asymptotic maximum likelihood theory suggests that

$$\hat{\underset{\sim}{\beta}} \overset{\cdot}{\sim} n_p(\underset{\sim}{\beta}, (X'\hat{W}X)^{-1})$$

where $\overset{\cdot}{\sim}$ denotes "approximately distributed for large N" and $n_p(\cdot,\cdot)$ denotes p variate normal with corresponding mean vector and covariance matrix. We assume throughout that appropriate conditions are satisfied for the above statement to hold.

In order to measure how well the LR model fits the data, Pregibon (1981) employs the "chi square" statistic

$$\chi^2 \equiv \chi^2(X\hat{\underset{\sim}{\beta}}) = (\underset{\sim}{y} - \hat{\underset{\sim}{y}})'\hat{W}^{-1}(\underset{\sim}{y} - \hat{\underset{\sim}{y}})$$

and the deviance

$$D \equiv D(X\hat{\underset{\sim}{\beta}}) = 2\{\ell(\hat{\underset{\sim}{\theta}}) - \ell(X\hat{\underset{\sim}{\beta}})\}$$

where $\hat{\underset{\sim}{\theta}}' = (\hat{\theta}_1,\ldots,\hat{\theta}_N)$, $p(\hat{\theta}_j) = y_j n_j^{-1}$, $(\hat{\theta}_i = \ln(y_j n_j^{-1}/(1 - y_j n_j^{-1})))$ $j = 1,\ldots,N$. It is natural to consider the χ^2 statistic after "thinking binomial". The Deviance measures how much "worse" one does by fitting the LR model than by fitting each observation separately. "Large" values for χ^2 and D indicate a poor fit. Asymptotically, χ^2 and D will be $\chi^2(N-p)$ (chi square with $N-p$

degrees of freedom) random variables under standard assumptions.

In the normal theory regression setting, uniquely defined residuals are employed to measure the fit of individual observations. It is clear from the above that there at least two natural ways to define residuals within the LR context. Define

$$\chi_j = (y_j - \hat{y}_j)/\hat{w}_j^{1/2}$$

and

$$d_j = \pm \sqrt{2}\ \{\ell_j(\hat{\theta}_j) - \ell_j(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}})\}$$

where $\ell_j(\hat{\theta}_j)$ is the j^{th} contribution to $\ell(\hat{\underset{\sim}{\theta}})$ and plus or minus is determined by whether or not $\hat{\theta}_j > \underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}$, (or equivalently according to whether or not $y_j n_j^{-1} > p_j$). The magnitudes of χ_i and d_i indicate how well case i fits the data and the signs indicate whether there is an overfit or underfit. The residuals d_i and χ_i are discussed in Pregibon (1981), while these and one other definition are discussed in Cook and Weisberg (1982). For general definitions of residuals, see Cox and Snell (1968).

It will be useful to consider Kullback-Leibler divergences, Kullback (1968) applied to Bernoulli distributions. Let $Z_i \sim \text{Bernoulli}(q_i)$, $i = 1,2$. Then the directed divergence between the distributions for Z_1 and Z_2 is defined as

$$I(q_1,q_2) = q_1 \ln q_1/q_2 + (1 - q_1) \ln (1 - q_1)/(1 - q_2),$$

which is non-negative definite, and is zero if and only if $q_1 = q_2$. It follows that the deviance may be expressed as

$$D = 2 \sum_{j=1}^{N} n_i\, I(y_j n_j^{-1}, \hat{p}_j)$$

and that

$$d_j = \pm\sqrt{2n_j}\ I^{1/2}(y_j n_j^{-1}, \hat{p}_j).$$

Thus the j^{th} contribution to the deviance d_j^2 is a measure of the discrepancy between the observed proportion of successes and that estimated by the LR model. The deviance itself is a weighted sum of these discrepancies, and is small when observed proportions match estimated proportions, and is large otherwise.

We now define the projection matrix

$$\hat{V} = \hat{W}^{1/2} X'(X\ \hat{W}X)^{-1}X'\hat{W}^{1/2} \equiv (\hat{v}_{ij})$$

and the residuals vector

$$\chi' = (\chi_1, \ldots, \chi_N).$$

Pregibon (1981) has shown that $\hat{V}\chi = 0$ which implies that the column space of V is orthogonal to the residuals vector. By analogy with the normal theory case, Pregibon suggests that large diagonals of $\hat{V}$ will correspond to extreme points in the design space, and hence are potentially influential. In the normal theory setting, Hoaglin and Welsch (1978) have called these "leverage" points and Johnson and Geisser (1983) have called them "distantly observed" points. Pregibon has also noted that pairs of observations with large values for $|\hat{v}_{ij}|$ will imply large effects of each case on the fit of the other. (See (3.2.2) and (3.4.6) in conjunction with (3.3.2).)

In order to get a clearer picture of the role of the matrix $\hat{V}$, consider the scaled vectors $\underset{\sim}{z}_i = \hat{w}_i^{1/2}\underset{\sim}{x}_i$ and the corresponding matrix $Z = \hat{W}^{1/2}X$. Since $\hat{V} = z(z'z)^{-1}z'$, it is possible to represent $\hat{V}$ in a way that leads to an interpretation of the diagonal components $(\hat{v}_{jj})$ as "distances", and the

off diagonal components ($\hat{v}_{ij}$) as measures of relative orientation of the scaled vectors $\underset{\sim}{z}_i$ and $\underset{\sim}{z}_j$. Let $\underset{\sim}{x}_j = (1, \underset{\sim}{\tilde{x}}_j)$, $\underset{\sim}{\tilde{z}}_j = \hat{w}_j^{1/2} \underset{\sim}{\tilde{x}}_j$, $\bar{\underset{\sim}{z}} = \sum_{j=1}^{N} \underset{\sim}{\tilde{z}}_j / N$, $S = \sum_{j=1}^{N} (\underset{\sim}{\tilde{z}}_j - \bar{\underset{\sim}{z}})^1 (\underset{\sim}{\tilde{z}}_j - \bar{\underset{\sim}{z}})/N$, and $Z = (z_1', \ldots, z_N')'$. Then

$$V = N^{-1}(\underset{\sim}{e}\,\underset{\sim}{e}' + (\tilde{Z} - \underset{\sim}{e}\,\bar{\underset{\sim}{z}})\, S^{-1}\, (\tilde{Z} - \underset{\sim}{e}\,\bar{\underset{\sim}{z}})')$$

where $\underset{\sim}{e}$ is an $N \times 1$ vector of ones. In particular

$$\hat{v}_{ij} = N^{-1}(1 + (\underset{\sim}{\tilde{z}}_j - \bar{\underset{\sim}{z}})\, S^{-1}\, (\underset{\sim}{\tilde{z}}_j - \bar{\underset{\sim}{z}})'). \tag{2.1.1}$$

Cases with large $\hat{v}_{jj}$ are "distantly observed" in the space spanned by the weighted covariate vectors $\{\underset{\sim}{\tilde{z}}_i\}_{j=1}^{N}$. It is to be noted that the weights depend heavily on the vectors $\{\underset{\sim}{\tilde{x}}_j\}_{j=1}^{N}$ since a "distant" $\underset{\sim}{\tilde{x}}_j$ will imply a small weight $\hat{w}_j$ (provided n_j is small). When all weights are similar in magnitude, the components of $\hat{v}_{ij}$ have the same interpretation as they do in the normal theory setting.

2.2 Influence

The topic of influence has been frequently discussed, most often within the context of normal theory regression. The recent monograph of Cook and Weisberg (1982) summarizes much of the recent work, including settings other than the normal.

Pregibon (1981) has discussed measures of influence which correspond to the effect an observation has on

(i) the estimation of $\underset{\sim}{\beta}$

(ii) the overall fit of the data to the LR model

(iii) the maximized likelihood

(iv) the fit of another observation (what he calls neighboring effects)

The particular quantities associated with these goals are

(i) $\hat{\beta}$,

(ii) $D(X\hat{\beta})$ and $\chi^2(X\hat{\beta})$,

(iii) $\ell(X\hat{\beta})$,

(iv) $\{d_i^2\}_{i=1}^N$.

The standard method of determining influence relative to a specific goal is to measure the effect particular observations have on the corresponding quantity of interest. If the goal is estimation, $\hat{\beta}$ is computed followed by computation of $\hat{\beta}_{(i)}$, the maximum likelihood estimate based on the full data set minus case i (subsequently referred to as the estimate based on retained data). We require some measure of the discrepancy (possibly a metric) between the two vectors, say $\hat{d}(\hat{\beta},\hat{\beta}_{(i)})$. Cases are then ordered according to the magnitudes of $\hat{d}(\hat{\beta},\hat{\beta}_{(i)})$. The paradigm is the same for other quantities, hence influence will depend not only on one's goals but also on one's choice of discrepancy measure. Measures specific to these goals are discussed in section 3.

We also consider measures of influence which correspond to the effect observations have on

(iv) the determination of probabilities

(v) the classification of observations into populations Π_1 and Π_0

(vi) each other relative to (iv) and (v)

(vii) the number of correct classifications in the sample

To fix notation let $X^f = (x_1^{f\prime},\ldots,x_m^{f\prime})'$ denote a set of covariates for M individuals who have not yet been classified as Π_0 or Π_1. (We use the superscript f to denote "future" observations.) When the goal is the determination of probabilities, we focus on

(iv) $\hat{p}^f \equiv p(X^f\hat{\beta})$

where we employ an obvious vector notation. When the goal is to classify

observations, we focus on the log odds ratio vector

(v) $X^f\hat{\underset{\sim}{\beta}} = \ln \hat{\underset{\sim}{p}}^f/(1-\hat{\underset{\sim}{p}}^f)$

since future j is classified as Π_0 or Π_1 according as $\ln \hat{p}_j^f/(1-\hat{p}_j^f) = \underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}$ is less than zero or greater than zero. We determine neighboring effects by focusing on the quantities

(vi) $\hat{p}_j^f$ and $\underset{\sim}{\hat{x}}_j^f\hat{\underset{\sim}{\beta}} = \ln \hat{p}_j^f/(1-\hat{p}_j^f)$, $\quad j = 1,\ldots,M,$

i.e. we will measure the effect of say case i on the determination of $\hat{p}_j^f$ and $\ln \hat{p}_j^f/(1-\hat{p}_j^f)$. When X^f is unavailable, we select $X^f = X$ to focus on the effects observations have on probabilities and classifications for the observed sample. We will finally consider a measure of the effect on the number of correct classification in the sample by focusing on

(vii) $$\text{NCCl}\ (X\hat{\underset{\sim}{\beta}}) = \sum_{j=1}^{N} y_j\, I_{(0,\infty)}(x_j\hat{\underset{\sim}{\beta}})$$

and

$$\text{NCCO}\ (X\hat{\underset{\sim}{\beta}}) = \sum_{j=1}^{N} (n_j - y_j)\, I_{(-\infty,0)}\ (x_j\hat{\underset{\sim}{\beta}}),$$

the estimated number of correct classifications into Π_1 and Π_0 respectively.

2.3 One Step Approximations

It is to be noted that all the quantities mentioned in section (2.2) depend on $\hat{\underset{\sim}{\beta}}$ and consequently it will always be necessary to compute $\hat{\underset{\sim}{\beta}}_{(i)}$, the estimate based on retained data. (The notation (i) will always imply that case i has been removed before calculation). Since it is necessary to iterate to obtain estimates of $\hat{\underset{\sim}{\beta}}_{(i)}$, it can be expensive to provide appropriate diagnostics. This is not a problem in the normal theory setting since $\hat{\underset{\sim}{\beta}}_{(i)}$ may be simply obtained from standard regression output c.f. Cook (1977), Cook and Weisberg (1980, 1982). Pregibon (1981) considers a "one step" approximation to $\hat{\underset{\sim}{\beta}}_{(i)}$ which he employs to obtain approximate

influence measures. Define $\ell(X_{(i)}\hat{\underset{\sim}{\beta}}_{(i)})$, the maximized log likelihood based on retained data and corresponding vector of first partials $\dot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}}_{(i)})(=\underset{\sim}{0})$ and matrix of second partials $\ddot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}}_{(i)})$. Then by first order Taylor expansion about $\hat{\underset{\sim}{\beta}}$, we obtain

$$\dot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}}_{(i)}) \doteq \dot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}}) + \ddot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}})(\hat{\underset{\sim}{\beta}}_{(i)} - \hat{\underset{\sim}{\beta}}),$$

from which we obtain the one step approximation

$$\hat{\underset{\sim}{\beta}}^{1}_{(i)} = \hat{\underset{\sim}{\beta}} - (-\ddot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}}))^{-1}\dot{\ell}(X_{(i)}\hat{\underset{\sim}{\beta}})$$

which has been shown by Cook and Weisberg (1982) to reduce to

$$\hat{\underset{\sim}{\beta}}^{1}_{(i)} = \hat{\underset{\sim}{\beta}} - (X'\hat{W}X)^{-1}\underset{\sim}{x}_i'(y_i - \hat{y}_i)/(1 - \hat{v}_{ii}). \tag{2.3.1}$$

Of course this approximation is best when the contours of the log likelihood are nearly elliptical, and can be very bad when this is not the case.

We will require one step approximations to the estimated probability vector $\hat{\underset{\sim}{p}}_{(i)} = \underset{\sim}{p}(X\hat{\underset{\sim}{\beta}}_{(i)}) = (\hat{p}_{1(i)},\ldots,\hat{p}_{N(i)})'$, and to the vector of log odds vector $X\hat{\underset{\sim}{\beta}}_{(i)} = \ln \hat{\underset{\sim}{p}}_{(i)}/(1 - \hat{\underset{\sim}{p}}_{(i)})$. Define the matrices

$$\Lambda = ({}_i\lambda_j) = (\underset{\sim}{x}_j(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)})), \quad \Lambda^1 = ({}_i\lambda^1_j) = (\underset{\sim}{x}_j(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}^1_{(i)})).$$

Then

$$\begin{aligned}\hat{p}_{j(i)} &= \exp(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}_{(i)})/(1 + \exp(\underset{\sim}{x}_j\underset{\sim}{\beta}_{(i)})) \\ &= \hat{p}_j \exp(-{}_i\lambda_j)/(1 - \hat{p}_j(1 - \exp(-{}_i\lambda_j))),\end{aligned}$$

and $\hat{p}^1_{j(i)}$ is defined by replacing ${}_i\lambda_j$ by ${}_i\lambda^1_j$ (equivalently define $\hat{\underset{\sim}{p}}^1_{(i)}$ as $\hat{p}(X\hat{\underset{\sim}{\beta}}^1_{(i)})$). The superscript "1" will always denote a one step approximation to the indicated quantity. We note for future reference that

$${}_i\lambda^1_j = \hat{w}_j^{-1/2}\chi_i\hat{v}_{ij}/(1 - \hat{v}_{ii}) \tag{2.3.2}$$

We similarly define $\Lambda^f = ({}_i\Lambda^f_j) = x^f_j(\hat{\beta} - \hat{\beta}_{(i)})$, $\Lambda^{fl} = ({}_i\lambda^{fl}_i)$ and $\hat{p}^{fl}_{(i)}$ for the case of future observations.

2.4 Calibration

Observations will be ordered from least to most influential according to a number of different measures. Having identified the "most" influential observation according to some measure, it is of interest to determine the order of magnitude of influence relative to other observations. We will refer to this as the problem of calibration. It will not be possible to calibrate all measures.

We follow a general procedure outlined by Cook (1977) and Cook and Weisberg (1982). Given an influence measure of the form $\hat{d}(\hat{\beta},\hat{\beta}_{(i)})$, we consider $d(\hat{\beta},\beta)$. Employing asymptotic maximum likelihood theory in conjunction with the delta method, it will sometimes be possible to identify the large sample distribution of $\hat{d}(\hat{\beta},\beta)$. It is then possible to equate the value $\hat{d}(\hat{\beta},\hat{\beta}_{(i)})$ to a percentage point of this distribution, and to use this for the purposes of comparison with other percentage points corresponding to other cases. For example if case i corresponds to the 50^{th} percentile and case j to the 90^{th} percentile of the same distribution, we have a better idea about how much more influential case j is than case i. It is important to point out that it is not appropriate to use this procedure as a method of testing whether or not observations belong to the assumed model.

3.0 Measures of influence

3.1 Estimation of β

Cook and Weisberg (1982) have defined a general measure of influence relative to the estimation of β. Define the log likelihood ratio

$$L(\hat{\beta},\beta) = 2(\ell(X\hat{\beta}) - \ell(X\beta)) \sim \chi^2(p) .$$

Then define the "likelihood distance"

$$LD_i = L(\hat{\underset{\sim}{\beta}}, \hat{\underset{\sim}{\beta}}_{(i)}). \tag{3.1.1}$$

We note that $LD_i \geq 0$ with equality when $X\hat{\underset{\sim}{\beta}} = X\hat{\underset{\sim}{\beta}}_{(i)}$, hence $\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)}$ need not equal $\underset{\sim}{0}$ for LD_i to be zero. The quantity LD_i measures the effect that the i^{th} case has on the maximized likelihood (insofar as it is affected by the change in $\hat{\underset{\sim}{\beta}}$), and may be calibrated by referring to the $\chi^2(p)$ distribution. A useful result is obtained by considering a second order expansion of $L(\hat{\underset{\sim}{\beta}}, \hat{\underset{\sim}{\beta}}_{(i)})$ about $\hat{\underset{\sim}{\beta}}$. It follows that

$$LD_i \doteq (\hat{\beta} - \hat{\beta}_{(i)})'(X'\hat{W}X)(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)}) \equiv D_i, \tag{3.1.2}$$

which is the natural adaptation of Cook's normal theory procedure referred to in section 1 c.f. Cook (1977, 1979). The approximate LD_i is seen to measure the global effect that case i has on the estimation of $\underset{\sim}{\beta}$ (relative to the covariability in $\hat{\underset{\sim}{\beta}}$). A one step approximation can be made to the approximate LD_i;

$$D_i^1 = \chi_i^2 \hat{v}_{ii}/(1 - \hat{v}_{ii})^2, \tag{3.1.3}$$

by application of (2.3.1) to (3.1.2). We note that cases which exhibit "lack of fit" (as measured by χ_i^2) and "distance" (as measured by $\hat{v}_{ii}$) will be most influential relative to the goal of estimating β, according to D_i^1. Appropriateness of the approximation, D_i^1, depends heavily on the shape of the likelihood. For further discussion and other measures, see Pregibon (1981) and Cook and Weisberg (1982).

3.2 Goodness of Fit

Assuming the goal is to measure the effect an observation has on the fit of the data to the LR model, we begin by noting that the likelihood distance may be interpreted as a goodness of fit diagnostic. Recalling the definition of $\hat{p}_{(i)}$, the likelihood distance may be expressed as

$$(3.2.1) \qquad LD_i = 2 \sum_{j=1}^{N} n_j \{ I(y_j n_j^{-1}, \hat{p}_{j(i)}) - I(y_j n_j^{-1}, \hat{p}_j) \}$$

$$= \sum_{j=1}^{n} (d^2_{j(i)} - d^2_j) \equiv \sum_{j=1}^{N} \Delta_i d_j$$

where $d^2_{j(i)}$ is defined in an obvious way. We note that $\Delta_i d_j$ measures the effect of the i^{th} observation on the fit of the i^{th} observation. When $\Delta_i d_j$ is positive (negative) the fit for the j^{th} observation is worse (better) after deletion. Pregibon (1981) obtained a one step approximation (after expansion) to $\Delta_i d_j$;

$$(3.2.2) \qquad \Delta_i d_j \doteq 2\chi_j \chi_i \hat{v}_{ij}/(1 - \hat{v}_{ij}) + \chi_i^2 \hat{v}_{ij}^2/(1 - \hat{v}_{ij})^2 \equiv {}_iD^1_j$$

He employs these to measure neighboring effects relative to fit. He also noted that

$$(3.2.3) \qquad \sum_{j \neq i} {}_iD^1_j \leq 0, \qquad {}_iD^1_i \geq 0$$

which implies that the overall fit for the retained observations must get better while that for case i must get worse after deletion of case i. We assume these results hold for the exact $\Delta_i d_j$ as well, and conclude that when LD_i is "large" it is due to an "appreciably" worse fit to the data after deletion, and most particularly, it will be due to a poor fit for case i. We may interpret LD_i to be the "overall" improvement in fit after deletion of case i since $LD_i \geq 0$.

We define the one step approximations LD_i^1 and $\Delta_i d_j^1$ to LD_i and $\Delta_i d_j$ respectively by substituting $\hat{p}_{j(i)}^1$ for $\hat{p}_j, j = 1,\ldots,N$. These approximations are made without expansion of LD_i and $\Delta_i d_j$. When the contours of the likelihood are nearly elliptical, $LD_i \doteq LD_i^1 \doteq D_i^1$ and $\Delta_i d_j \doteq \Delta_i d_j^1 \doteq {}_iD_j^1$, while otherwise they may be appreciably different. We would expect that the approximations LD_i^1 and $\Delta_i d_j^1$ to be better than D_i^1 and ${}_iD_j^1$ since fewer approximations are involved. This is not generally the case however as will be seen in the examples.

Pregibon (1981) considered the deviance effect measure

$$\Delta_i D = D(X\hat{\underset{\sim}{\beta}}) - D_{(i)}(X_{(i)}\hat{\underset{\sim}{\beta}}_{(i)})$$

which is asymptotically $\chi^2(1)$ and thus is easily calibrated. This measure may be expressed as

$$\Delta_i D = d_i^2 - \sum_{j\neq i} \Delta_i d_j = d_i^2 - LD_i + \Delta_i d_i$$

and thus may be interpreted by (3.2.1) and (3.2.3) as the lack (or goodness) of fit of the i^{th} observation plus the overall improvement in fit of the retained data. The second expression indicates the relationship between LD_i and $\Delta_i D$. Pregibon obtains the one step approximation (after expansion)

$$\Delta_i D^1 \equiv d_i^2 + \chi_i^2 \cdot \hat{v}_{ii}/(1-\hat{v}_{ii}) \tag{3.2.4}$$

which implies that the overall improvement in fit for retained data is approximately

$$\sum_{j\neq i} {}_iD_j^1 = -\chi_i^2 \hat{v}_{ii}/(1-\hat{v}_{ii}).$$

We close this section by noting another result of Pregibon (1981). He shows that

(3.2.5) $$\Delta_i\chi^2 = \chi^2 - \chi^2_{(i)} \doteq \chi^2_i/(1-\hat{v}_{ii})$$

after approximation. This is analogous to the squared studentized residual in regression c.f. Behnken and Praper (1972), and measures the effect of case i on the overall fit of the model to the data. Cases which are distant in the weighted design space are thus potentially more influential relative to fit.

3.3 Determination of Probabilities

Since the primary goal of any analysis of this sort is essentially prediction, it is of interest to measure the differential effects of observations on the determination of $\underset{\sim}{p}$ and $\underset{\sim}{p}^f$, a vector of probabilities associated with future or, as yet, unclassified individuals. It is even more compelling that we focus on these aspects in view of the fact that most influence measures considered in this paper already depend on the difference between $\hat{\underset{\sim}{p}}$ and $\hat{\underset{\sim}{p}}_{(i)}$.

We begin by supposing that interest is focused on the determination of the $\underset{\sim}{p}^f = \underset{\sim}{p}(X^f\underset{\sim}{\beta})$. A somewhat naive measure of the influence that case i has on the entire vector $\hat{\underset{\sim}{p}}^f$ is the Euclidean distance between the vectors $\hat{\underset{\sim}{p}}^f = \underset{\sim}{p}(X^f\hat{\underset{\sim}{\beta}})$ and $\hat{\underset{\sim}{p}}^f_{(i)} = \underset{\sim}{p}(X^f\hat{\underset{\sim}{\beta}}_{(i)})$,

(3.3.1) $$ED_i^f = \{(\hat{\underset{\sim}{p}}^f - \hat{\underset{\sim}{p}}^f_{(i)})'(\hat{\underset{\sim}{p}}^f - \hat{\underset{\sim}{p}}^f_{(i)})\}^{1/2} = (\Sigma_j(\hat{p}^f_j - \hat{p}^f_{j(i)})^2)^{1/2}$$

A problem with this measure is illustrated by supposing $\hat{p}^f_j = .001$, $\hat{p}^f_{j(i)} = .01$, and $\hat{p}^f_k = .501$, $\hat{p}^f_{k(i)} = .51$. Then the contribution to ED_i^f from the j^{th} and k^{th} observations will be the same while it is clear that case i is much more influential in it's effect on the determination of p^f_j than it is on the determination of p^f_k. To alleviate this problem, we could consider a different inner product, say $\hat{L} \equiv \text{diag}\{\ell_1,\ldots,\ell_m\}$. Then define $\hat{\underset{\sim}{\ell}} = (\hat{\ell}_1,\ldots,\hat{\ell}_m)'$ and $ED_i^f(\hat{\underset{\sim}{\ell}}) = \{(\hat{\underset{\sim}{p}}^f - \hat{\underset{\sim}{p}}^f_{(i)})'L(\hat{\underset{\sim}{p}}^f - \hat{\underset{\sim}{p}}^f_{(i)})\}^{1/2}$. The proper choice of $\hat{L}$ is not

obvious, except for the fact that in general it must depend on the data, and that in examples like the one above, $\hat{\ell}_j$ should be larger than $\hat{\ell}_k$. If we assume $Y_i^f \sim \text{Bin}\ (m_i, p_i^f)$, the choice $\hat{\underset{\sim}{\ell}} = \underset{\sim}{m} = (m_1, \ldots, m_M)'$ results in the diagnostic

$$ED_i^f(\underset{\sim}{m}) = \{(\underset{\sim}{y}^f - \underset{\sim(i)}{y}^f)'(\underset{\sim}{y}^f - \underset{\sim(i)}{y}^f)\}^{1/2}$$

the Euclidean distance between the vectors of predictions based on full and retained data respectively.

While it is not possible to calibrate $ED_i^f(\hat{\underset{\sim}{\ell}})$, it is possible to calibrate the individual effects measures ${}_ie_j^f \equiv p_j^f - p_{j(i)}^f$, $j = 1, \ldots, M$. Since

$$\hat{p}_j^f - p_j^f \dot{\sim} n(0,\ \hat{m}_{jj}^f \{\hat{p}_j^f (1 - \hat{p}_j^f)\}^2)$$

where $\hat{M}^f = (\hat{m}_{ij}^f) = (\underset{\sim i}{x}^f (X'\hat{W}X)^{-1} \underset{\sim j}{x}^{f\prime})$, it follows that

$$(\hat{p}_j^f - p_j^f)/(\hat{m}_{jj}^f)^{\frac{1}{2}} \hat{p}_j (1 - \hat{p}_j) \dot{\sim} n(0,1), \tag{3.3.2}$$

and thus we may employ the one step diagnostic

$${}_i\tilde{e}_j^{f1} \equiv (\hat{p}_j^f - \hat{p}_{j(i)}^{f1})/(\hat{m}_{jj}^f)^{\frac{1}{2}} \hat{p}_j (1 - \hat{p}_j) \tag{3.3.3}$$

as a standardized measure of the effect that case i has on the determination of $\hat{p}_j^f$. We may refer ${}_i\tilde{e}_j^{f1}$ to the n(0,1) distribution for the purposes of calibration. Since (3.3.2) may be employed to obtain a large sample confidence interval for p_j, we can suggest that removal of case i will result in an estimate of p_j which is moved to the edge of a confidence interval with confidence coefficient corresponding to the normal percentile for ${}_i\tilde{e}_j$.

As an alternative to Euclidean distance, we consider the symmetric Kullback-Leibler divergence, c.f. Kullback (1968), as a discrepancy measure between probabilities. In order to measure the effect case i in the sample

has on future case j, define

$$ {}_i g_j^f = I(\hat{p}_j^f,\ \hat{p}_{j(i)}^f) + I(\hat{p}_{j(i)}^f,\ \hat{p}_j^f). $$

Then to measure the collective effect case i has the entire future sample, define

$$ \text{DIV}_i^f = \sum_{j=1}^{m} {}_i g_j^f. \tag{3.3.4} $$

We note that the difficulty exhibited in the previous example has been lessened since ${}_i g_j^f = .0023$ and ${}_i g_k^f = .000036$. We generalize the definition of influence here to include the possibility of further weighting of the individual effects. Let $\hat{L}$ and $\hat{\underset{\sim}{\ell}}$ be defined as above. Then

$$ \text{DIV}_i^f(\hat{\underset{\sim}{\ell}}) = \sum_{j=1}^{m} \hat{\ell}_j ({}_i g_j^f), $$

and simple calculation results in

$$ \begin{aligned} \text{DIV}_i^f(\hat{\underset{\sim}{\ell}}) &= \sum_{j=1}^{m} \hat{\ell}_j (\hat{p}_j^f - \hat{p}_{j(i)}^f) \ln \hat{p}_j^f (1-\hat{p}_j^f)^{-1} / \hat{p}_{j(i)}^f (1-\hat{p}_{j(i)}^f)^{-1} \\ &= \sum_{j=1}^{m} \hat{\ell}_j (\hat{p}_j^f - \hat{p}_{j(i)}^f)\ \underset{\sim}{x}_j^f (\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)}) \\ &= \sum_{j=1}^{m} \hat{\ell}_j (\hat{p}_j^f - \hat{p}_{j(i)}^f)({}_i \lambda_j^f). \end{aligned} $$

We note that $\text{DIV}_i^f(\hat{\underset{\sim}{\ell}})$ weights each difference in estimated probabilities with a log relative odds ratio. The one step approximation $\text{DIV}_i^{f1}(\hat{\underset{\sim}{\ell}})$ is defined by substituting ${}_i \lambda_j^{f1}$ and $\hat{p}_{j(i)}^{f1}$ for ${}_i \lambda_j^f$ and $\hat{p}_{j(i)}^f$ respectively.

In the absence of a clear choice for $\hat{\underset{\sim}{\ell}}$, we will let $\hat{\underset{\sim}{\ell}} = \underset{\sim}{e}$, the vector of ones. However we keep the more general notation for comparative purposes.

It is not possible to calibrate $\text{DIV}_i^f(\hat{\ell})$ in general. Further, it is not always possible to pre-specify X^f. When this is the case, it is reasonable to choose $X^f = X$ and to determine the effects cases have on $\hat{\underset{\sim}{p}}$, the vector

of probabilities corresponding to the observed sample. (See Johnson and Geisser (1983) for examles in the normal theory setting.) Measures are defined as above only with the superscript f deleted. In this case we obtain an interesting result if we let $\hat{\ell}_j = n_j$, $j = 1,\ldots,N$. Define

$$S_j(\hat{\underset{\sim}{\beta}},\beta) = n_j(\hat{p}_j - p_j)\{\underset{\sim}{x}_j(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})\},$$

and expand about $\hat{\underset{\sim}{\beta}}$, to obtain

$$S_j(\hat{\underset{\sim}{\beta}},\underset{\sim}{\beta}) \doteq \hat{w}_j\{\underset{\sim}{x}_j(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})\}^2$$

and hence

$$DIV_i(\underset{\sim}{n}) \doteq \sum_{j=1}^{N} \hat{w}_j({}_i\lambda_j)^2.$$

Recalling the approximation to the likelihood distance, we obtain

$$LD_i \doteq (\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)})'X'\hat{W}X(\hat{\beta} - \hat{\beta}_{(i)}) = \sum_{j=1}^{N} \hat{w}_j({}_i\lambda_j)^2 \doteq DIV_i.$$

Thus, the divergence which weights each case with the number of individuals corresponding to that case, and the likelihood distance, may be approximated by each other. Accordingly, we have a new interpretation for LD_i as an approximate measure of the discrepancy between probabilities estimated before and after deletion. This also provides a justification for calibration of $DIV_i^1(\underset{\sim}{n})$ by referring to the $\chi^2(p)$ distribution.

We observe in passing that the weighted distance $ED_i(\underset{\sim}{n}^2)$ may be similarly approximated as

$$ED_i(\underset{\sim}{n}^2) = \left(\sum_{j=1}^{N} w_j^2({}_i\lambda_j)^2\right)^{\frac{1}{2}}$$

where $\underset{\sim}{n}^2 \equiv (n_1^2,\ldots,n_N^2)'$.

It is informative to take another look at LD_i from a different perspective. We note that

$$(3.3.5) \qquad \exp(LD_i) = \prod_{j=1}^{N} p(Y_j = y_j | \underset{\sim}{x}_j, \hat{\underset{\sim}{\beta}}) / p(Y_j = y_j | \underset{\sim}{x}_j, \hat{\underset{\sim}{\beta}}_{(i)})$$

which is a likelihood ratio statistic. We may think of the probability functions $p(Y_j = y_j | x_j, \hat{\underset{\sim}{\beta}})$ and $p(Y_j = y_j | x_j, \hat{\underset{\sim}{\beta}}_{(i)})$ as predictive distributions and the ratio (3.3.4) as the relative odds of observing the data actually observed under repeated predictive trials from the two distributions above, for given X. The likelihood distance LD_i may thus be interpreted as a measure of the effect deletion of case i has on the classical joint predictive distribution (evaluated at observed data)

$$p(\underset{\sim}{y} | X, \hat{\underset{\sim}{\beta}}) \equiv \prod_{j=1}^{N} p(Y_j = y_j | \underset{\sim}{x}_j, \hat{\underset{\sim}{\beta}}).$$

Johnson and Geisser (1983) determine the effects observations have on joint Bayesian predictive distributions in the normal theory setting. This approach is not so easily adapted here due to technical difficulties. We note however that if we employ the approximation based on retained data $p(Y | X, \hat{\underset{\sim}{\beta}}_{(i)})$, then the Kullback-Leibler symmetric divergence between these two approximate joint predictive distributions is exactly DIV_i. Thus we have further justification for DIV_i, as well as LD_i, as measures of the effect case i has on the prediction of observations or equivalently, on the determination of probabilities.

We finally note that Larimore (1983) has discusses the likelihood ratio (3.3.5) and has essentially discussed a large sample version of the result (3.2.1) within the context of model selection.

3.4 Classification

Of course the determination of probabilities and the classification of observations are highly related processes since observations are classified according to the magnitudes of corresponding probabilities. Observations which affect estimated probabilities should also be influential regarding classification of observations, and vice versa. However there is extra information to be gained by further characterizing influence according to the effects observations have on classification.

We assume the same setup as in (3.3), and we suppose that cases will classified according to the rule

(3.4.1) classify future case j as Π_1 (Π_0) if $\underset{\sim}{x}_j^f \hat{\underset{\sim}{\beta}} = \ln \hat{p}_j^f/(1-\hat{p}_j^f) > 0$ (< 0).

The quantity $x_j^f \hat{\underset{\sim}{\beta}}$ is analogous to Fisher's discriminant function in the normal theory case, and the rule above corresponds to equal losses for both types of classification error.

In order to assess the effect that observations have on this rule, we define the log odds measure

$$(3.4.2) \qquad LO_i^f(\hat{\underset{\sim}{\ell}}) = \sum_{j=1}^{M} \hat{\ell}_j \ln \{\hat{p}_j^f(1-\hat{p}_j^f)^{-1}/\hat{p}_{j(i)}^f(1-\hat{p}_{j(i)}^f)^{-1}\}$$

$$= \sum_{j=1}^{M} \hat{\ell}_j \{\underset{\sim}{x}_j^f(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)})\} = \sum_{j=1}^{M} \hat{\ell}_j ({}_i\lambda_j^f)$$

which has one step approximation $LO_i^1(\hat{\underset{\sim}{\ell}})$ where ${}_i\lambda_j^{f1}$ has been substituted for ${}_i\lambda_j^f$. Since effects may cancel one another, we also consider the absolute log odds measure

$$(3.4.3) \qquad ALO_i^f(\hat{\underset{\sim}{\ell}}) = \sum_{j=1}^{M} \hat{\ell}_j |{}_i\lambda_j^f|$$

and corresponding one step approximation $ALO_i^{f1}(\hat{\ell})$, which may not be calibrated. We determine individual effects by considering ${}_i\lambda_j^f$ which, on the other hand, may be calibrated due to the fact that

$$x_j^f(\hat{\beta} - \beta) \sim n(0, \hat{m}_{jj}^f).$$

We define

$$(3.4.4) \qquad {}_i\tilde{\lambda}_j^{f1} = {}_i\lambda_j^{f1}/(\hat{m}_{jj}^f)^{1/2}$$

which may be referred to a n(0,1) distribution. Removal of case i results in moving estimated log odds $(x_j\hat{\beta})$ to the edge of a confidence interval with percent coverage determined by to the corresponding percentile for ${}_i\tilde{\lambda}_j^{f1}$.

We may also calibrate $LO_i^f(\underset{\sim}{e}) \equiv LO_i^f$ since

$$\sum_{j=1}^{M} \underset{\sim}{x}_j^f(\hat{\beta} - \beta) \sim n(0, \underset{\sim}{e}'\hat{M}^f\underset{\sim}{e}).$$

Define $\hat{\sigma}^f = (\underset{\sim}{e}'\hat{M}^f\underset{\sim}{e})^{1/2}$, and the standardized log odds measure (based on equal weights)

$$(3.4.5) \qquad SLO_i^{f1} = \sum_{j=1}^{M} {}_i\lambda_j^{f1}/\hat{\sigma}^f$$

which may be referred to the n(0,1) distribution. This measure will not be very useful when effects on different cases cancel due to differences in signs. However when SLO_i^f is "large" and positive (or negative), the implication is that the future samples are more likely to be allocated to Π_0 (Π_1) after deletion of case i than before; i.e. deletion of case i results in an overall decrease (increase) in the $\hat{p}_j^f$, j = 1,...,m, after deletion. Of course when ${}_i\lambda_j^f$ is "large" and positive or negative, the same statement applies to $\hat{p}_j^f$ alone. In those cases where effects cancel, the measure ALO_i^f may be employed as a "backup" measure. Individual effects may then be measured via ${}_i\lambda_j^{f1}$.

As before, when X^f is not available, we let $X^f = X$ and make definitions

as above with the superscript "f" deleted. We note some of the comparisons among influence measures when $X^f = X$:

$$
\begin{aligned}
&LO_i \doteq \sum_j \hat{w}_j ({}_i\lambda_j)^2 \doteq DIV_i,\ LO_i = \sum_j {}_i\lambda_j, \\
&ALO_i = \sum_j |{}_i\lambda_j|,\ ED_i \doteq \sum_j \{\hat{p}_j(1-\hat{p}_j)\}^2 ({}_i\lambda_j)^2.
\end{aligned} \tag{3.4.6}
$$

The approximate likelihood distance D_i, the divergence DIV_i, and the Euclidean distance ED_i, may thus be interpreted as weighted measures of the effect of case i on the squared log posterior odds. Recalling from (2.3.2) that the one step approximation to ${}_i\lambda_j$ may be expressed as

$${}_i\lambda_j^1 = \chi_i (1-\hat{v}_{ii})^{-1/2} \{\hat{v}_{ij}/\hat{v}_{ii}^{1/2}\ \hat{v}_{ij}^{1/2}\} \{\hat{v}_{ii}/(1-\hat{v}_{ii})\}^{1/2}\ \hat{m}_{jj}^{1/2},$$

we may further interpret one step approximations to the expressions in (3.4.6). It is convenient to note that

$${}_i\tilde{\lambda}_j^1 = \chi_i (1-\hat{v}_{ii})^{-1/2} \{\hat{v}_{ij}/\hat{v}_{jj}^{1/2}\ \hat{v}_{ii}^{1/2}\} \{\hat{v}_{ii}/(1-\hat{v}_{ii})\}^{1/2}$$

which is the product of the analogue to the Studentized residual referred to at (3.2.5), the estimated correlation, $\text{corr}(w_i^{1/2} \underset{\sim}{x}_i \hat{\beta}, w_i^{1/2} \underset{\sim}{x}_j \hat{\beta})$, and the "distance" measure $\{\hat{v}_{ii}/(1-\hat{v}_{ii})\}^{1/2}$. Hence lack of fit, "relative orientation", and "distance" (see the discussion following (2.1.1)) determine the magnitude and sign of ${}_i\tilde{\lambda}_j^1$ and hence ${}_i\lambda_j^1$, where orientation and distance refer to the weighted covariate vectors $\hat{w}_i^{1/2} \underset{\sim}{x}_i$ and $\hat{w}_j^{1/2} \underset{\sim}{x}_j$.

We conclude this section by defining quantities for determining the approximate influence observations have on the number of correctly classified observations in the sample. Define

$$NCCl_i = \sum_j y_j\{I_{(0,\infty)}(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}) - I_{(0,\infty)}(x_j\hat{\beta}_{(i)})\}$$

$$NCCO_i = \sum_j (n_j - y_j)\{I_{(-\infty,0)}(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}) - I_{(-\infty,0)}(\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}_{(i)})\}.$$

the differences in the number of correct classifications, into Π_1 and Π_0 respectively, before and after deletion of case i. We obtain one step approximations $NCCl_i^1$ and $NCCO_i^1$ by substituting $\underset{\sim}{x}_j\hat{\underset{\sim}{\beta}}^1_{(i)} = \underset{\sim}{x}_j\underset{\sim}{\beta} - {}_i\lambda_j^1$. When $NCC\ell_i^1$ is > (<) 0 the implication is that fewer (more) observations from Π_ℓ are correctly classified after deletion of case i than before, $\ell = 0,1$.

4. Examples

4.1 Leukemia Data

Feigel and Zelen (1965) analyzed a data set consisting of 33 observations on the survival of individuals diagnosed with leukemia (see figure 1). Observed covariates were WBC = white blood cell count and the variable AG which indicates the presence or absense of a certain morphologic characteristic in the white cells. Cook and Weisberg (1982) also analyzed this data set from the point of view of detecting influential observations. They define a "success" to correspond to patient survival in excess of 52 weeks, and "failure" otherwise. Note that there are 30 cases due to three multiplicities at (AG = 1, WBC = 100,000) and two at (AG = 0, WBC = 100,000). Since it is expected that individuals with high WBC are at high risk, it is clear that case 15 will be very influential since one of the five individuals with WBC = 100,000 apparently survived at least 52 weeks from diagnosis. It is to be noted that all 16 of the remaining individuals with WBC larger than 15,000 died within 52 weeks.

The top 5 influential cases and corresponding influence statistics are listed in Table 1. Case 15 stands out as most influential according to all measures, except for those related to the number of correct classifications. It is clear though that removal of case 15 will affect all inferential goals considered in this paper. Note that D_i^1 is a better approximation to LD_i than LD_i^1 is for this data set and that both approximations are best when LD_i is small. The approximation $\Delta_i D^1$ is very good for small to moderate $\Delta_i D$, however it is not as good for $\Delta_{15} D$.

The large positive value for SLO_{15} (may be referred to the 87^{th} percentile of the standard normal distribution) indicates that removal of case 15 will result in lower probabilities of success on the average and consequently an increased number of individuals will be allocated as failures. Careful consideration of figures 1 and 2 makes this fact clear. From figure 1 it can be seen that future individuals in the regions (WBC < 12,000 or WBC > 30,000, and AG = 1 or 0) and (AG = 0, 11,000 < WBC < 28,000) will be given the same allocation according to the discriminants $\underset{\sim}{x}_j^f\hat{\underset{\sim}{\beta}}$ and $\underset{\sim}{x}_j^f\hat{\underset{\sim}{\beta}}_{(15)}$, while those individuals in (AG = 1, 12,000 < WBC < 30,000) will be allocated as failures by $\underset{\sim}{x}_j^f\hat{\underset{\sim}{\beta}}_{(15)}$ and successes by $\underset{\sim}{x}_j^f\hat{\underset{\sim}{\beta}}$. Since only case 8 falls into the latter region, it follows that $NCCO_{15} = 1$, $NCC1_{15} = 0$, and hence the approximation $NCCO_{15}^1 = 0$ has failed in this instance. The index plots for ${}_{15}\lambda_j$ in figure 2 indicate that removal of case 15 results in a relatively large decrease in posterior odds (or equivalently a relatively large increase in probability of failure) for cases 12, 13, 14, 15, 29 and 30 and moderate decreases for case 24, 25, 26, 27 and 28.

We turn to a comparison of measures of the neighboring effects of case 15 on the determination of probabilities. We may compare the measures ${}_{15}\tilde{\lambda}_j^1$, ${}_{15}\tilde{e}_j^1$, and ${}_{15}g_j^1$ by consideration of figure 2. It is to be noted that while ${}_i\tilde{e}_j^1$ and ${}_i\tilde{\lambda}_j^1$ are asymptotically equivalent, they can be considerably different.

It is also to be noted that ${}_{15}g_j^1$ indicates relatively diminished effects for some cases. For example, the effect of case 15 on case 30 is large according to ${}_{15}\tilde{\lambda}_{30}$ (2.66), moderate according to ${}_{15}e_{30}^1$ (.56) (both may be referred to the standard normal distribution) and small according to ${}_{15}g_{30}^1$ (.05) (when compared to max ${}_{15}g_j^1 = .65$). In fact since $\hat{p}_{30} = .0112$ and $\hat{p}_{30(15)} = .0001$ ($\hat{p}_{30(15)} = 0.0000$), one's choice of measure will depend upon the emphasis one wishes to place on this type of discrepancy. In any event, all three measures agree that case 15 affects the determination of probabilities for cases 12, 13, 14 and 15 (all cases have high WBC and AG = 1; $\hat{p}_j - \hat{p}_{j(15)} = .26, .28, .28, .10$, respectively).

We finally discuss the effects case 15 has on the fit of other cases. Figure 3 contains index plots of ${}_{15}D_j^1/n_j$ and ${}_{15}\Delta d_j^1/n_j$. Recall that ${}_{15}\Delta d_j^1$ measures how much better or worse case j fits the model after removal of case i, and that ${}_{15}D_j^1$ is an approximation to ${}_{15}\Delta d_j^1$ (see (3.2.1), (3.2.2) and (3.2.3)). Case 15 has the greatest effect on itself, fitting much worse after deletion than before, while cases 12, 13, 14, 16 and 17 fit better. The approximations are very similar. As a quick check on accuracy we note that ${}_{15}\Delta d_{15}/3 = 5.61$, ${}_{15}\Delta d_{15}^1/3 = 3.09$ and ${}_{15}D_{15}^1/3 = 4.48$. Pregibon (1981) has discussed the accuracy of ${}_iD_j^1$ and found it to be quite good for the Finney data discussed in 4.2.

Removal of case 15 results in virtually no influential observations (see table 1). The accuracy of the approximations is again to be noted. For this data set, the approximation D_i^1 is to be preferred to LD_i^1.

4.2 Finney data

Finney (1947) studied the relationship between rate and volume of air inspired on a transient vaso-constriction of the skin of the digits. He defined a "success" to be the occurrence, and a "failure", the nonoccurrence of vaso-constriction. The data consists of thirty nine observations on three

individuals and is plotted in figure 4. Pregibon (1981) has carefully studied this data set from the point of view of detecting influential observations.

Influence statistics for this data are listed in table 2. Cases 4 and 18 are most influential according to the likelihood distance, divergence, absolute log odds, $\Delta_i\chi^2$, and $\Delta_i D$. Case 32 on the other hand is most influential according to Euclidean distance, standardized log odds, and has the greatest change in the number of correctly classified observations. Figure 4 contains plots of discriminant lines for models based on full and retained data (models based on data without cases 4 and 32, respectively). These lines are fairly similar. (Lines corresponding to deletion of 18, and the pair (4,18) are not drawn since they are virtually identical to the one computed without case 4.) Deletion of cases 4, 18, or 32 results in a modest clockwise shift of the discriminant line. When case 32 is deleted, three more failures are classified as successes than would be if the full data set was employed (note that $NCCO^1_{32} = -3$). When cases 4, 18 or 32 are deleted, standardized log odds statistics are relatively small due to a cancellation of effects. Deletion of case 4 or 18 does result in a slightly increased propensity to allocate cases as failures, and deletion of case 32 results in the increased propensity to allocate cases as successes, on the average. These phenomena are better understood after careful scrutiny of figure 4.

Allocation of future cases in the region between lines will be affected by one's choice of model. When lines move perpendicularly after deletion, the cancellation of effects does not occur, and this is indicated by the fact that ALO_i and LO_i are similar in magnitude. Since ALO_i appears large relative to LO_i for cases 4, 18, and 32, it will be necessary to consider the measures $_i\tilde{\lambda}_j$ (see figure 5) in order to determine the effects these cases have on other cases.

ALO_i and LO_i are similar in magnitude. Since ALO_i appears large relative to LO_i for cases 4, 18 and 32, it will be necessary to consider the measures ${}_i\tilde{\lambda}_j$ (see figure 5) in order to determine the effects these cases have on other cases.

The effect on the determination of probabilities is measured by the likelihood distance, divergence, and Euclidean distance. The approximations D_i^1 and LD_i^1 are reasonably good, and D_i^1 is again an improvement over LD_i^1. The values for LD_i, D_i^1, LD_i^1, and DIV_i^1 may all be referred to the $\chi^2(3)$ distribution for calibration. None of the cases appear to be exceedingly influential in this regard.

Index plots for ${}_4\tilde{\lambda}_j$ and ${}_{32}\tilde{\lambda}_j$ are given in figure 5. These indicate that cases 4 and 18 affect the determination of probabilities (or log odds) for many cases in a modest way (many values for ${}_4\tilde{\lambda}_j$ are near 1 which may be referred to the standard normal distribution). Most probabilities are decreased after deletion of case 4 or 18, while deletion of case 32 has less of an effect on the determination of nearly all probabilities. We note that the plot for ${}_{18}\tilde{\lambda}_j$ is nearly identical to that for ${}_4\tilde{\lambda}_j$ and the plots for ${}_4\tilde{e}_j$ and ${}_{32}\tilde{e}_j$ are nearly identical to those for ${}_4\tilde{\lambda}_j$ and ${}_{32}\tilde{\lambda}_j$ respectively, so they are not given.

Most cases are not greatly affected by cases 4, 18 or 32 according to the measure ${}_ig_j^1$. The largest effect is that which case 32 has on itself (${}_{32}g_{32}^1 = .18$). We find that $\hat{p}_{32} \doteq .42$ and $\hat{p}^1_{32(32)} \doteq .63$ ($p_{32(32)} \doteq .64$). Regarding the effects of cases 4 and 18, note that ${}_4g_4^1 = .04 = {}_{18}g_4^1$ and ${}_4g_{18}^1 = .05 = {}_{18}g_{18}^1$. We find that $\hat{p}_4 \doteq .07$, $\hat{p}^1_{4(4)} = \hat{p}^1_{4(18)} \doteq .03$, and that $\hat{p}_{18} = .08$, $\hat{p}^1_{18(18)} = \hat{p}^1_{18(4)} = .03$. As was the case with the leukemia data, the measure ${}_ig_j^1$ does not generally emphasize the same cases as ${}_i\tilde{\lambda}_j^1$.

Effects on fit are indicated by $\Delta_i D$. Cases 4 and 18 are clearly very influential in this regard. Individual effects are measured by ${}_iD_j^1$ and ${}_\Delta d_j^1$. Cases 4 and 18 may still be treated symmetrically. The fit of most

cases is not affected much by cases 4 or 18 according to these measures. However the magnitude of the effect cases 4 and 18 have on each other ranges from 1.75 to 2.05. Cases 4 and 18 are both "successes" and estimated probabilities decrease by .04 and .05 respectively, indicating the worse fit after deletion.

Removal of cases 4 and 18 simultaneously results in several very influential observations among those remaining (see table 2). The reason appears to be that after deletion, successes and failures are nearly perfectly divided by the line $\underset{\sim}{x}\hat{\beta}_{(4,18)} = 0$, and influential cases are those that fall on or nearly on that line. In fact, removal of case 39, a success just barely to the left of the line, results in the failure of the convergence algorithm for the maximum likelihood estimates.

4.3 Population data

Population change data were collected from census records for the fifty states of the U.S. by Press and Wilson (1978). The percent increase in total population of each state was noted and the median increase for all states determined. States were allocated as "successes" and "failures" according as their percent increase was above or below the median, respectively. Observed covariates were per capita income (INC), birth rate (BR), death rate (DR), urbanization of population (UR) and presence or absence of coastline (CO). See Press and Wilson (1978) for more details.

The determination of influential observations depends on the model which is selected. A standard LR analysis based on the full data set would result in the deletion of UR due to a small value for the asymptotic test statistic. However it is possible that an influential data point could be responsible for this small value. In fact, deletion of case 35

(Florida) results in an appreciable increase in the test statistic for inclusion of UR. We choose, however, to proceed with UR deleted from the model since deletion of the "most" influential case (Louisiana) does not effect an appreciable change in the test statistic. It can be seen from table 3 that Louisiana (case 12) is very influential according to all measures except $\Delta_i D$, both when UR is included and when it is deleted. Other states are decidedly less influential.

Louisiana has a relatively large likelihood distance ($LD_{12} = 5.46$) which may be referred to a $\chi^2(5)$ distribution. (Note that the approximations D^1_{15} and LD^1_{15} are too large in this instance, and that LD^1_{15} is an improvement over D^1_{12}). Louisiana is thus very influential in its effect on the determination of probabilities, the estimation of $\underset{\sim}{\beta}$, and is potentially influential regarding the classification of future observations. These conclusions are further supported by the fact that values for DIV^1_{12} (5.43), SLO^1_{12} (.91) and ALO^1_{12} (34.29) are relatively large. And since SLO^1_{12} is positive, the discriminant hyperplane will be moved, after deletion of Louisiana, in such a way that proportionately more future cases may be allocated as "failures." This possibility is supported by the data since $NCCO_{12} = 2$, and $NCCl_{12} = -2$ ($NCCO^1_{12} = 1$, $NCCl^1_{12} = -2$), which indicates that four additional cases will be classified as failures after deletion.

The effect of removing Louisiana on the determination of probabilities and log odds for other states (neighboring effects) can be noted by looking at figures 6 and 7. Louisiana has a large effect on many states including itself. The measures ${}_{12}\tilde{\lambda}_j$ and ${}_{12}\tilde{e}_j$ are nearly identical for most cases (cases 12, 23, 30, and 32 are the exceptions). The remarkable difference for case 32 can be explained by the fact that $\hat{p}_{32} = .99$ and

$\hat{p}^1_{32(12)} = .77$ which results in a relatively large numerator and a small denominator for ${}_{12}\tilde{e}_{32}$ (see (3.3.3)). Cases with small or large probabilities before deletion are weighted more heavily by ${}_i\tilde{e}_j$ than by ${}_i\tilde{\lambda}_j$ (see (3.3.3) and (3.4.4)). Figure 7 gives the actual (one step) differences in probabilities. With the exception of case 32, the information obtained from figures 5 and 6 appears to be qualitatively the same for this data set.

The most influential case relative to fit is New York ($\Delta_{37}D = 4.64$). This case is not very influential relative to any other criterion, and so we proceed to determine effects for Louisiana ($\Delta_{12}D = 3.29$). We note that Louisiana has a large effect on the fit of itself ($\Delta_{12}d^1_{12} = 8.16$), and a moderate effect on the fit of New York ($\Delta_{12}d^1_{37} = -.87$) and Wisconsin ($\Delta_{12}d^1_{39} = -.89$). Since $\hat{p}_{12} \doteq .55$, $\hat{p}^1_{12(12)} \doteq .01$ and Louisiana's population increase was above the median, Louisiana clearly fits much worse after deletion. Similarly note that $\hat{p}_{39} \doteq .30$, $\hat{p}_{39(12)} \doteq .47$, $\hat{p}_{37} \doteq .85$ and $\hat{p}_{37(12)} \doteq .77$, and that Wisconsin's population increase was above, and New York's was below the median increase, which indicates the fact that these states fit somewhat better after deletion.

Deletion of Louisiana results in accepting a model without a constant and without the variable UR. It can be seen from table 3 that there are no observations which are particularly influential under these circumstances.

4.4 Diagnosis data

As a final example, we briefly consider a data set consisting of only 21 observations on differential diagnoses of Cushing's syndrome. The data is studied in Aitchison and Dunsmore (1975, p. 212). There are three types of syndrome; adenoma, bilateral hyperplasia, and carcinoma. We combine adenoma and bilateral hyperplasia into one group which will be referred to as "failures", and individuals with carcinoma will be termed "successes". Covariates are taken to be the natural logs of urinary excretion rates of two steroid metabolites which we identify as ln T and ln P. The data is plotted in figure 8, and influence statistics are given in table 4.

It is clear from table 4 that one step approximations to LD_i are not adequate for this data. Cases 12 and 19 are extremely influential in that their removal results in perfect separation of "successes" and "failures" (see figure 8), and consequently in a nearly perfect fit of the data to the model. Case 19 is detected as most influential according to ALO_i and SLO_i. However, since $ALO_{19} \approx LO_{19} > 0$, the implication is that nearly all cases have lower probability of success after deletion. This is not the case since the probabilities for failures are decreased (to zero) those for successes are increased (to one) after deletion. These measures have also not detected case 12. Cases 12 and 19 are detected, however, as most influential according to $\Delta_i D^1$ and $\Delta_i D$, the change in deviance measures, as well as the measure $\Delta_i \chi^2$. The one step approximations are better here than for other measures.

A problem with this data is that it takes 19 iterations to get convergence of the maximum likelihood algorithm when cases 12 or 19 are removed and only 8 iterations when case 1 is removed. Thus one step approximations may be too far away from the fully iterated values to be appropriate.

In any event, it is clear that removal of cases which result in better separation of the data are potentially very influential regarding the determination of probabilities, and consequently, the classification of observations, as well as the fit of the model. It is also clear that these cases can be difficult to detect, especially as the dimension p increases, unless proper care is taken.

5.0 Concluding Remarks

In this paper we have discussed some existing measures of influence, and have proposed and discussed some new measures. We must conclude that measures which detect influence relative to the determination of probabilities are the most relevant and useful. The magnitudes of all measures discussed are highly dependent upon the effect observations have on $\hat{\underset{\sim}{p}}$. Of course it could be argued that it is really the influence on $\hat{\beta}$ that should be focused on since it is the effect on $\hat{\beta}$ that determines the effect on $\hat{\underset{\sim}{p}}$. However it seems much more appealing to focus on probabilities for individuals rather than artificially constructed regression coefficients vectors, and it is easier to compare the components of the vectors $\hat{\underset{\sim}{p}}$ and $\hat{\underset{\sim}{p}}_{(i)}$ than it is to compare components of $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}_{(i)}$, since probabilities are more easily interpreted and since they are restricted to the unit interval.

For detecting influence, we prefer the measure DIV_i^1 because of it's justification as a measure of the effect on the joint estimative predictive distribution for a future sample with the same covariates. However it has been noted that $DIV_i^1 \doteq D_i^1 \doteq LD_i^1 \dot{\sim} \chi^2_{(p)}$, when $n_1 = n_2 = \ldots = n_N = 1$, and that D_i^1 appears to be a better approximation to LD_i than LD_i^1 is. When the n_i's are not all ones, DIV_i^1 and D_i^1 will differ due to the differential weighting of cases for D_i^1. In any case, D_i^1 may still be interpreted as a weighted mea-

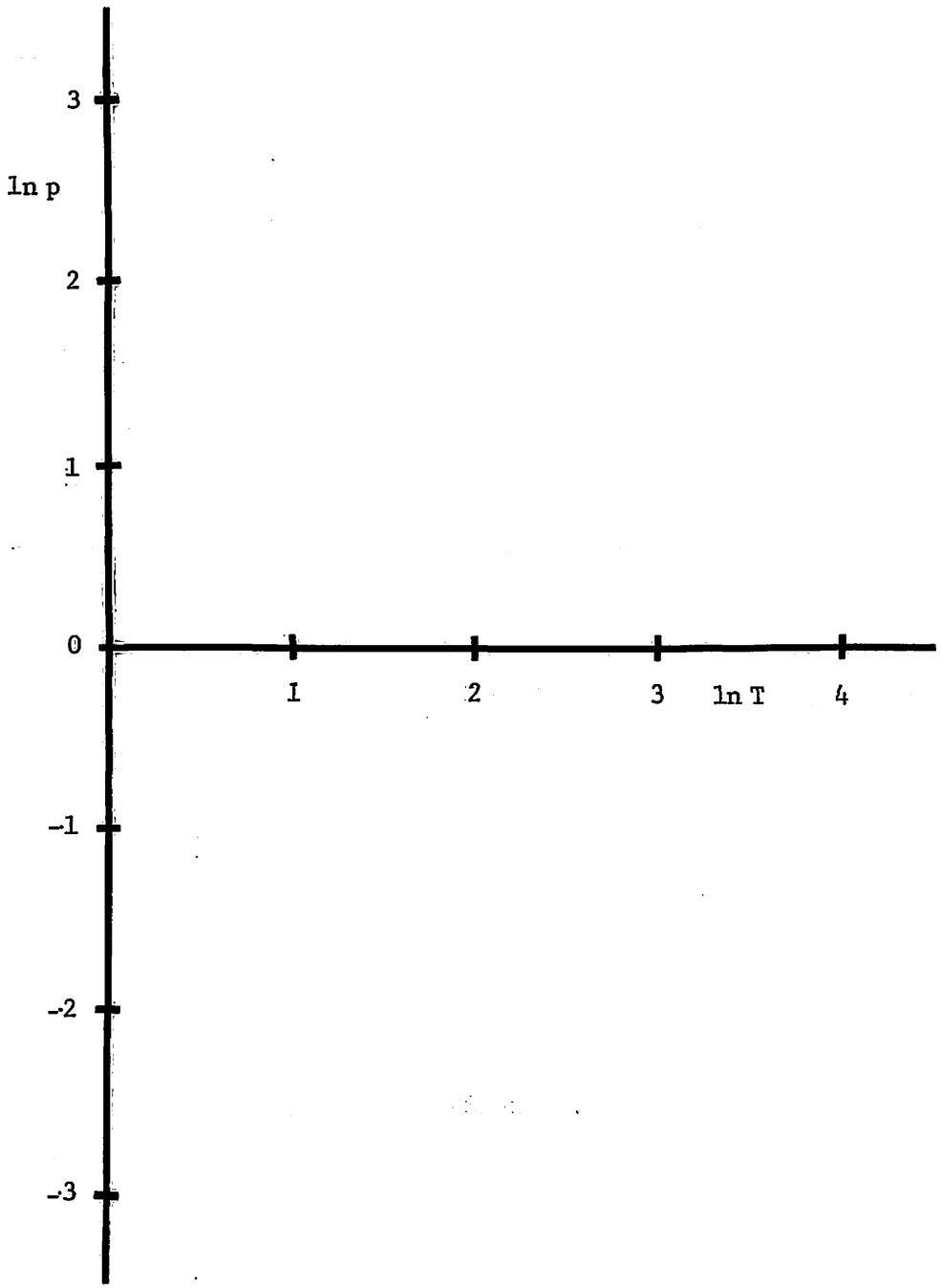
sure of the effect cases have on the determination of probabilities, and since it is less expensive to compute than DIV_i^1, some may prefer to use it. We note, however, that for the samples considered in this paper, the difference in expense was not great.

Measures relative to the fit of observations and the classification of observations are useful for detecting influence. However, once it is known that a case affects $\hat{p}$, it is also essentially known that the fit will be affected and that classifications will be affected. The summary measures LO_i^1, ALO_i^1, SLO_i^1, $NCCl_i$, $NCCO_i$ and $\Delta_i D^1$ are thus recommended as secondary measures for the purpose of characterizing the influence of observations that have already been detected by DIV_i^1 (or D_i^1).

For the purpose of further characterizing influence, we recommend calculation of ${}_i\tilde{\lambda}_j^1$, ${}_i\tilde{e}_j^1$, ${}_ie_j^1$, ${}_ig_j^1$, and $\Delta_i d_j^1$ (or ${}_iD_j^1$) (or some subset chosen according to preferance), for cases already identified as influential. In this way, one can determine precisely the influence exerted on the sample by the identified observation.

As a final remark, we caution that where deletion of cases results in "near" separation of the data by a hyperplane, one step approximations may fail to adequately detect influence.

Acknowledgement: I would like to thank Shen-Yen Lin for computational assistance.


3
ln p
2
1
0
-1
-2
-3
1
2
3
ln T
4

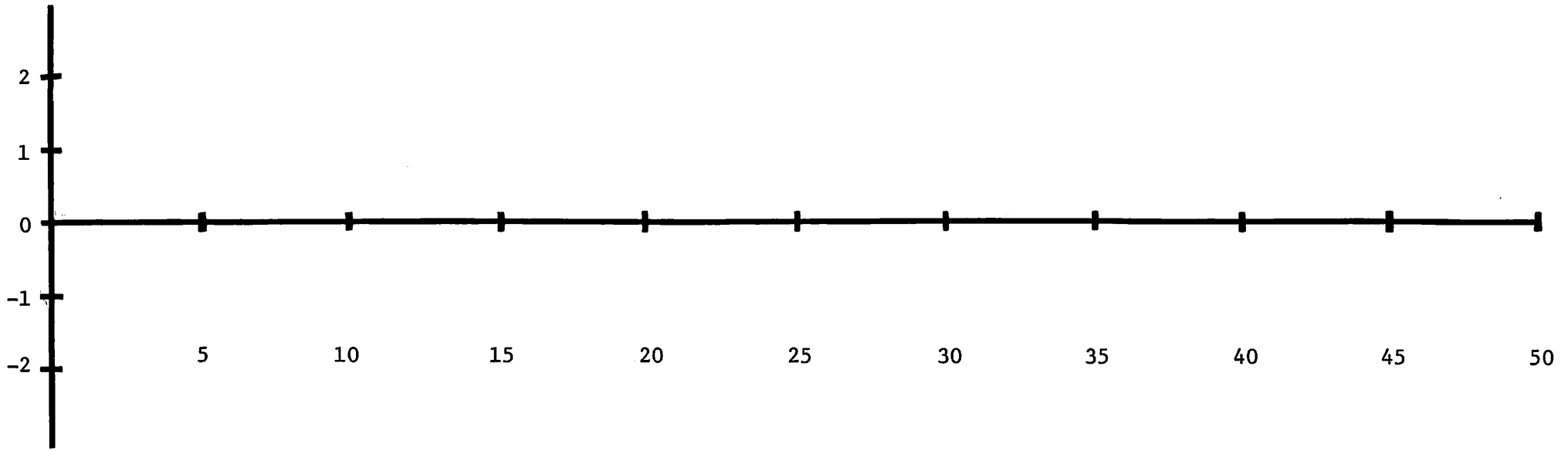


Figure 6.

Index plots for population data (fall data without UR) corresponding to $\{_{12}\tilde{\lambda}_j\}$ (solid line) and $\{_{12}\tilde{e}_j\}$ (dashed line).

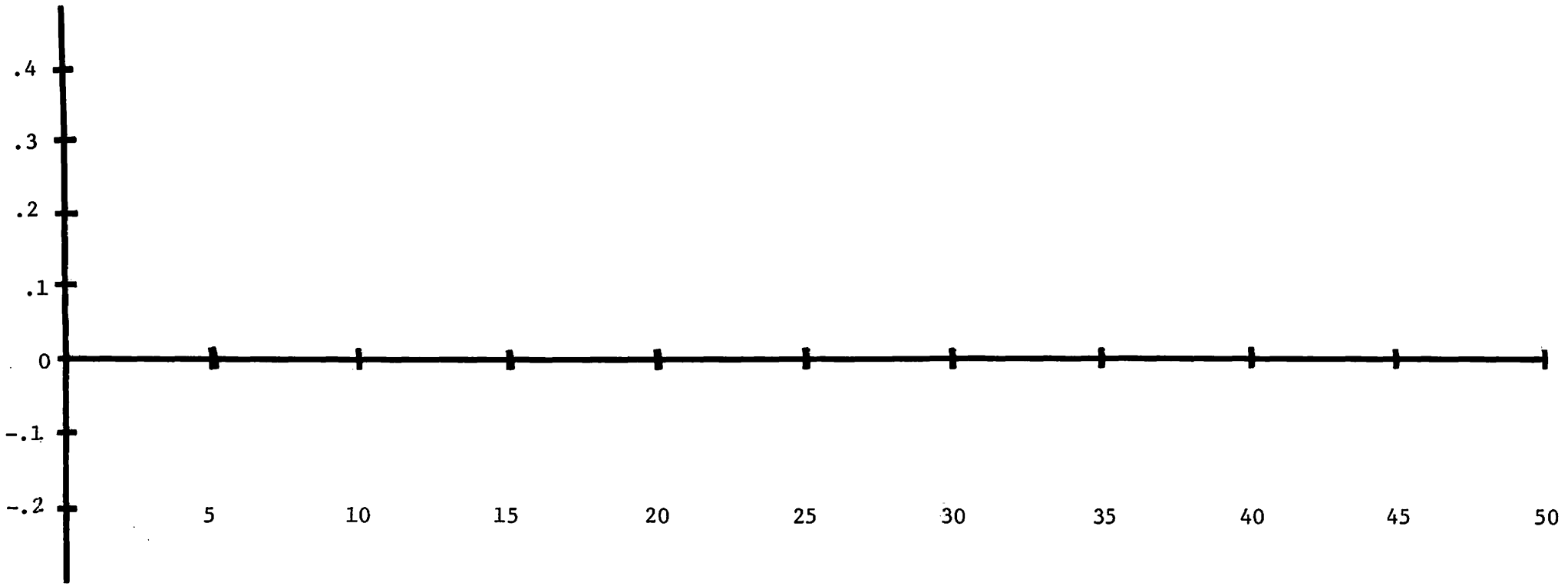


Figure 7.

Index plot for population data full data without UR, corresponding to $\{_{12}e_j\}$.

BIBLIOGRAPHY

Aitchison, J. and Cook, R.D. (1977). "Detection of influential observations in linear regression". Technometrics 19, 15-18.

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press. Cambridge.

Cook, R.D. (1979). "Influential observations in linear regression". Journal of the American Statistical Association 74, 169-174.

Cook, R.D. and Weisberg, S. (1980). "Characterizations of an empirical influence function for detecting influential cases in regression". Technometrics 22, 495-508.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York. Chapman and Hall.

Cox, D.R. and Snell, E.J. (1968). "A general definition of residuals". Journal of the Royal Statistical Society. Series B, 30, 248-275.

Feigel, P. and Zelen, M. (1965). "Estimation of exponential probabilities with concomitant information". Biometrics 21, 826-838.

Finney, D.J. (1947). "The estimation from individual records of the relationship between dose and quantal response". Biometrika 34, 320-334.

Hoaglin, D.C. and Welsch, R. (1978). "The hat matrix in regression and ANOVA". American Statistician 32, 17-22.

Johnson, W. and Geisser, S. (1981). "On the architecture of estimative influence functions". University of California, Davis Technical Report #23.

Johnson, W. and Geisser, S. (1982). "Assessing the predictive influence of observations". Statistics and Probability: Essays in Honor of C.R. Rao. 343-358.

Johnson, W. and Geisser, S. (1983). "A predictive view of the detection and characterization of influential observations in regression". Journal of the American Statistical Association 78, 137-144.

Kullback, S. (1968). Information Theory and Statistics. Peter Smith. Gloucester, Massachusetts.

Larimore, W.E. (1983). "Predictive inference, sufficiency, entropy, and an asymptotic likelihood principle". Biometrika 70, 175-181.

Nelder, J. and Wedderburn, R. (1972). "Generalized linear model". Journal of the Royal Statistical Society. Series A, 135, 370-384.

Pregibon, D. (1981). "Logistic regression diagnostics". Annals of Statistics. 9, 705-724.

Press, S.J. and Wilson, S. (1978). "Choosing between logistic regression and discriminant analysis". Journal of the American Statistical Association. 73, 699-705.

Table 1. Influence measures for leukemia data

Case #	D_i^1	LD_i^1	LD_i	DIV_i^1	ED_i^1	LO_i^1	SLO_i^1	ALO_i^1	$\Delta_i X^*$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCI_i^1$	$\hat{v}_{ii}$	Data set
15	9.94	4.85	10.72	2.67	.59	16.27	1.13	28.07	2.32	4.74	8.14	0	0	.65	Leukemia (Full)
17	.51	.45	.71	.41	.22	7.50	.52	8.42	2.13	3.70	3.77	0	0	.10	
16	.50	.45	.68	.41	.22	7.58	.53	8.40	2.17	3.77	3.84	0	0	.10	
9	.21	.21	.24	.20	.20	-2.77	-.19	4.75	-1.54	2.51	2.52	-2	0	.08	
5	.21	.21	.22	.19	.20	-2.77	-.19	4.65	-1.53	2.48	2.49	-2	0	.08	
9	.52	.47	.67	.44	.25	5.79	.12	17.57	-2.17	3.79	3.85	0	0	.11	Leukemia (without case 15, without constant)
5	.47	.43	.57	.41	.25	3.09	.06	14.42	-2.04	3.54	3.59	0	0	.11	
8	.16	.14	.13	.18	.16	-16.78	-.35	16.91	-.67	.69	.68	0	0	.34	
16	.12	.11	.16	.10	.10	16.17	.21	18.33	1.66	2.69	2.71	0	0	.19	
7	.12	.12	.12	.12	.15	7.22	.15	7.22	.87	1.10	1.10	0	0	.15	

* $\Delta_i X = X_i/(1-\hat{v}_{ii})^{\frac{1}{2}}$

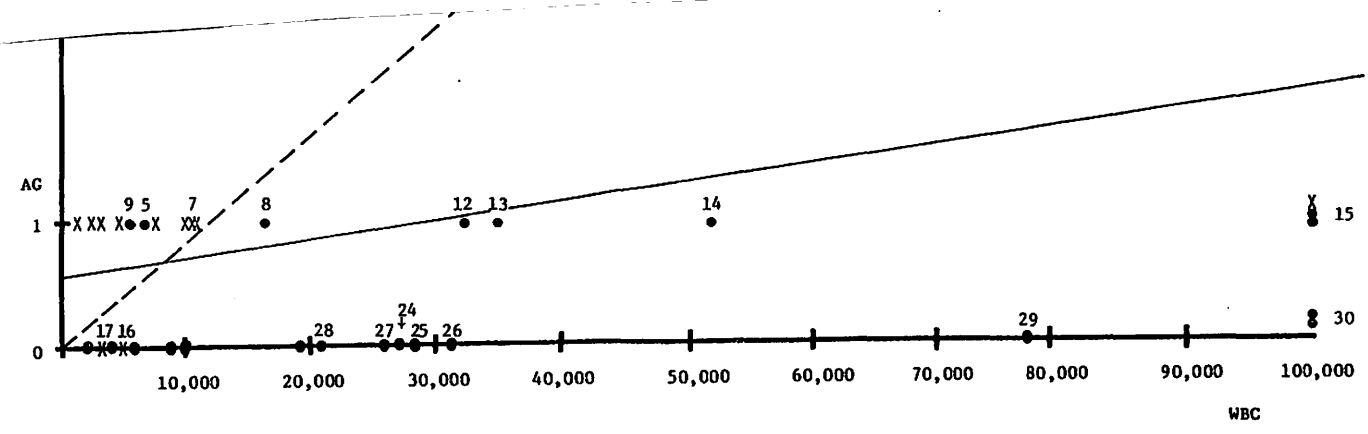


Figure 1

Leukemia Data: (X) indicates "success" and (•) indicates failure. The lines (—) and (---) satisfy $\underset{\sim}{x}\hat{\underset{\sim}{\beta}} = 0$ and $\underset{\sim}{x}\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)} = 0$ where $\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)}$ is the regression coefficients vector determined without case 15 and without a constant. Values below the respective lines are allocated a "failures".

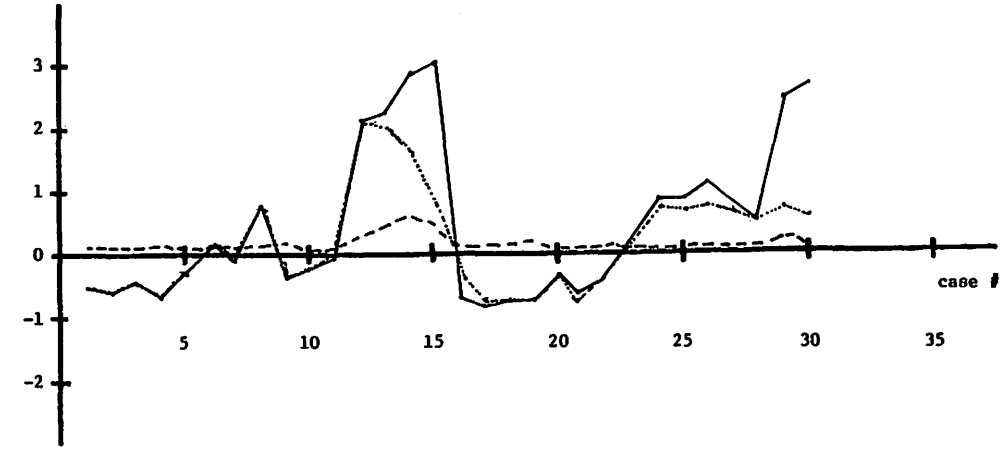


Figure 2.

Index plots for leukemia data corresponding to: $\{_{15}\tilde{\lambda}_j^1\}$ (solid line), $\{_{15}\tilde{e}_j^1\}$ (dotted line) and $\{_{15}g_j^1\}$ (dashed line).

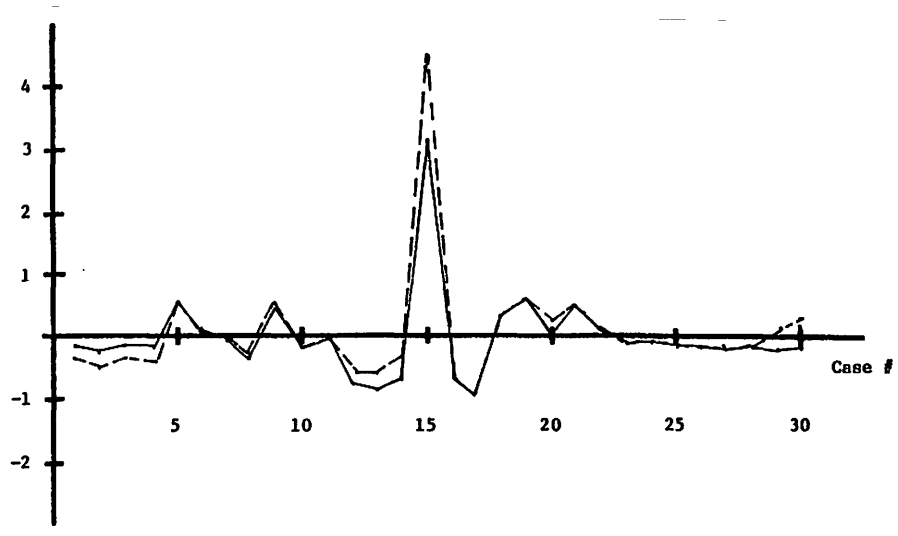


Figure 3.

Index plots for leukemia data corresponding to: $\{\Delta_{15}d_j^1\}$ (solid line) and $\{{}_{15}D_j^1/n_j\}$ (dashed line).

Table 2. Influence statistics for Finney's data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	AlO_i^1	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
4	1.05	.86	1.59	.79	.30	4.95	.27	28.90	3.88	6.40	6.70	0	0	.07	Finney (Full)
18	.91	.76	1.47	.69	.28	5.29	.29	26.58	3.58	5.96	6.19	0	0	.07	
32	.55	.55	.58	.54	.33	-7.06	-.39	12.25	-1.04	1.45	1.45	-3	0	.33	
13	.49	.45	.59	.44	.27	-5.29	-.29	10.04	-1.59	2.68	2.71	-1	0	.16	
12	.20	.20	.22	.20	.19	-3.30	-.18	8.31	-1.03	1.45	1.64	-1	0	.16	
13	3.59	2.70	8.17	2.34	.55	-39.84	-1.21	100.74	-2.00	4.15	4.97	0	0	.47	Finney (without cases 4 and 18)
32	1.68	1.69	1.97	1.62	.58	-15.39	-.47	28.45	-1.14	1.64	1.68	-1	0	.56	
39	1.30	.90	*	.77	.27	8.03	.24	163.82	2.66	4.98	*	0	0	.16	
35	.18	.19	.18	.20	.20	13.38	.41	34.05	.87	1.10	1.09	0	0	.19	
34	.16	.17	.16	.18	.19	12.16	.37	40.96	.73	.81	.10	0	0	.23	

* indicates that estimates would not converge.

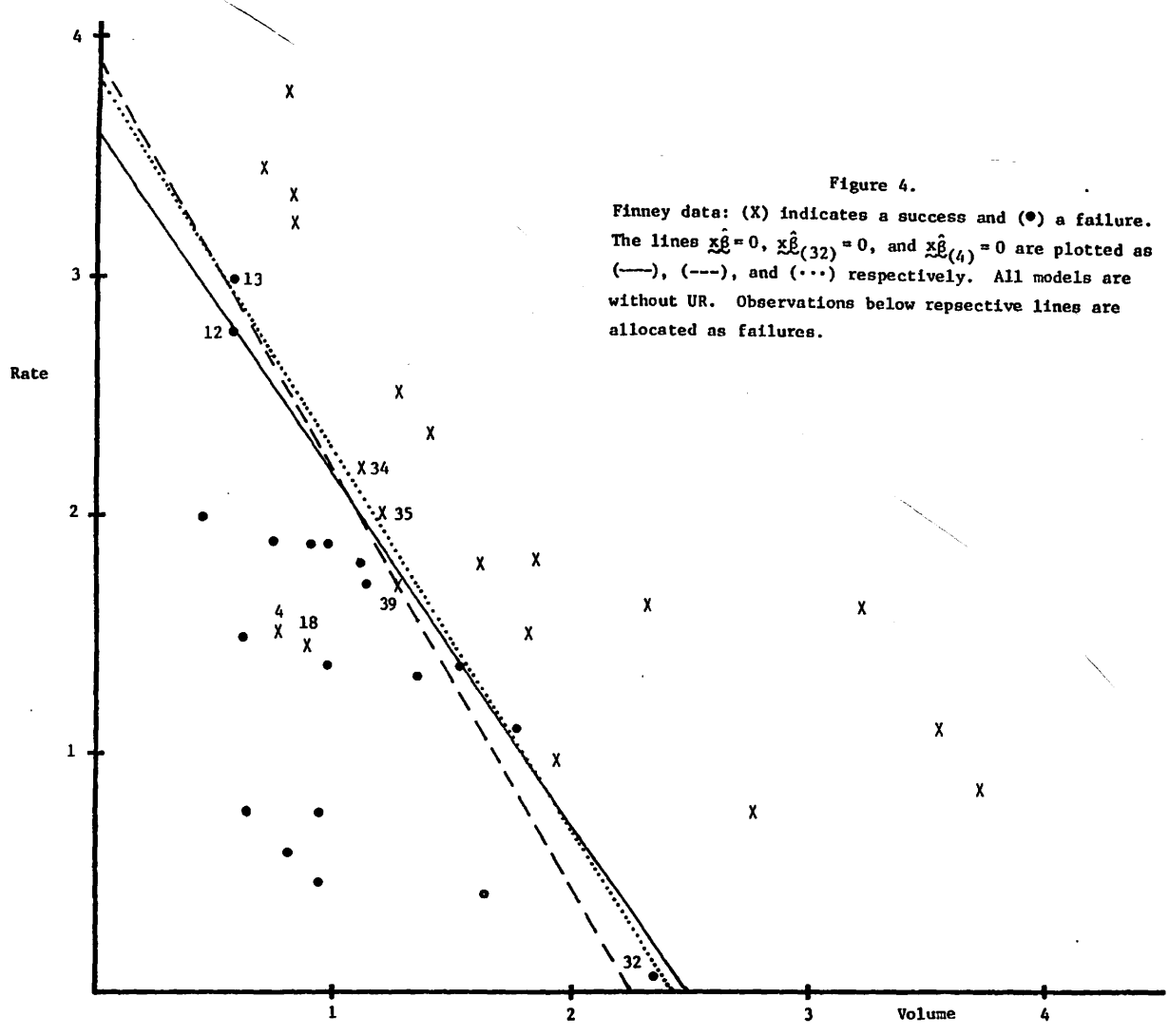


Figure 4.

Finney data: (X) indicates a success and (•) a failure. The lines $\underset{\sim}{x}\hat{\underset{\sim}{\beta}} = 0$, $\underset{\sim}{x}\hat{\underset{\sim}{\beta}}_{(32)} = 0$, and $\underset{\sim}{x}\hat{\underset{\sim}{\beta}}_{(4)} = 0$ are plotted as (——), (---), and (···) respectively. All models are without UR. Observations below repsective lines are allocated as failures.

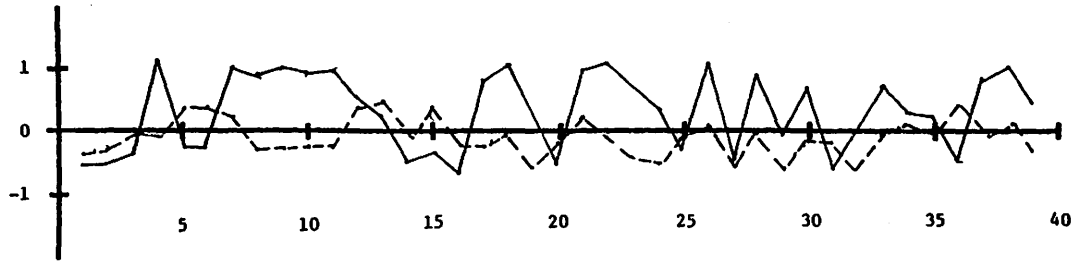


Figure 5.

Index plots for Finney's data corresponding to $\{{}_4\tilde{\lambda}_j\}$ (solid line) and $\{{}_{32}\tilde{\lambda}_j\}$ (dashed line).

Table 3. Influence statistics for population data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	AlO_i^1	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
12	7.56	6.41	5.61	5.54	.84	17.63	.93	33.14	1.53	2.68	2.63	1	-3	.76	Population (Full)
35	1.17	1.03	1.44	.97	.39	8.56	.45	23.61	2.57	4.77	4.90	1	-1	.15	
47	1.25	1.11	1.45	1.05	.42	2.67	-.14	21.71	-2.11	3.98	4.08	-1	0	.22	
37	1.10	.97	1.41	.91	.37	-9.16	-.48	21.38	-2.83	5.13	5.27	0	0	.12	
9	.43	.40	.51	.38	.23	6.86	.36	16.02	2.62	4.43	4.47	0	-2	.06	
12	8.06	6.27	5.46	5.48	.83	17.10	.91	34.29	1.73	3.38	3.29	1	-2	.73	Population (without UR)
47	1.26	1.13	1.46	1.07	.43	3.36	-.18	21.58	-2.07	3.91	3.99	-1	1	.23	
35	.43	.40	.45	.39	.26	6.74	.36	13.13	1.84	3.16	3.18	1	0	.11	
9	.42	.39	.48	.37	.22	6.76	.36	15.76	2.71	4.55	4.59	0	0	.05	
37	.89	.81	1.07	.77	.34	16.86	-.44	16.86	-2.53	4.56	4.64	0	0	.12	
47	1.03	.90	1.33	.84	.36	-3.43	-.18	18.68	-2.11	3.90	4.03	-1	3	.11	Population (without constant, UR, or case 12)
37	.73	.66	.90	.63	.30	-6.77	-.35	17.30	-2.31	4.12	4.20	0	0	.12	
35	.57	.52	.66	.50	.28	9.15	.47	15.96	2.22	3.89	3.94	1	0	.11	
9	.45	.41	.54	.40	.23	8.40	.44	17.68	2.85	4.76	4.80	1	0	.05	
24	.34	.33	.36	.32	.22	-6.03	-.31	10.06	-1.61	2.68	2.70	0	0	.09	

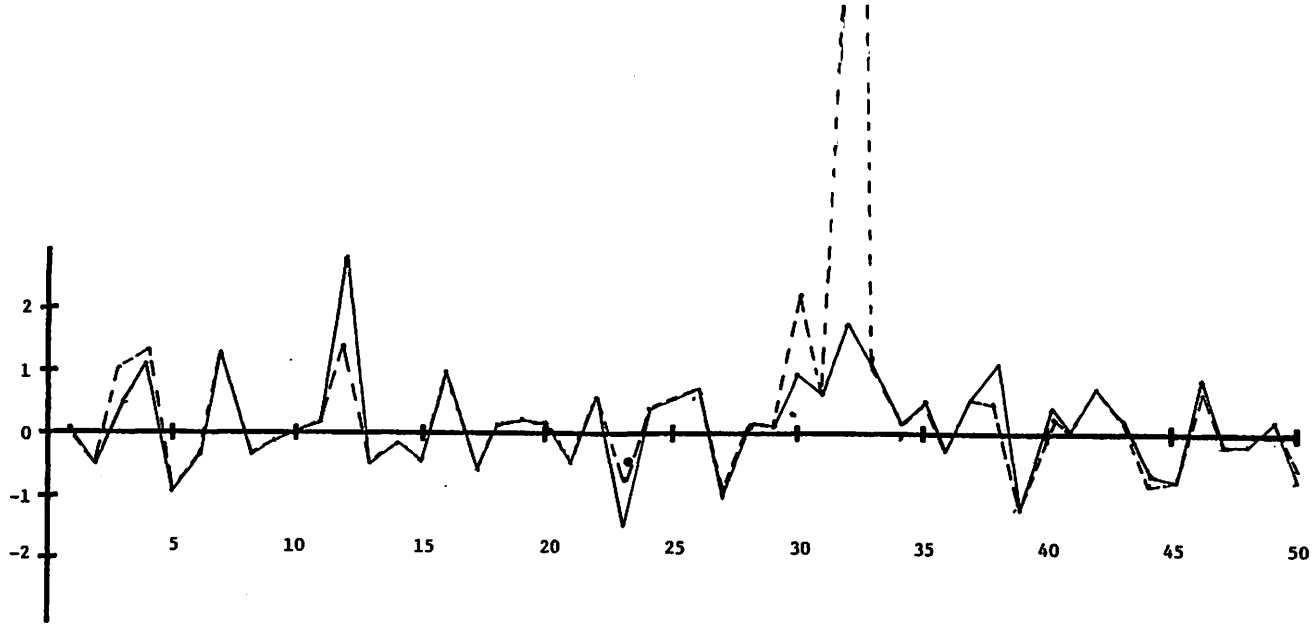


Figure 6.
Index plots for population data (fall data without UR) corresponding to $\{_{12}\tilde{\lambda}_j\}$ (solid line) and $\{_{12}\tilde{e}_j\}$ (dashed line).

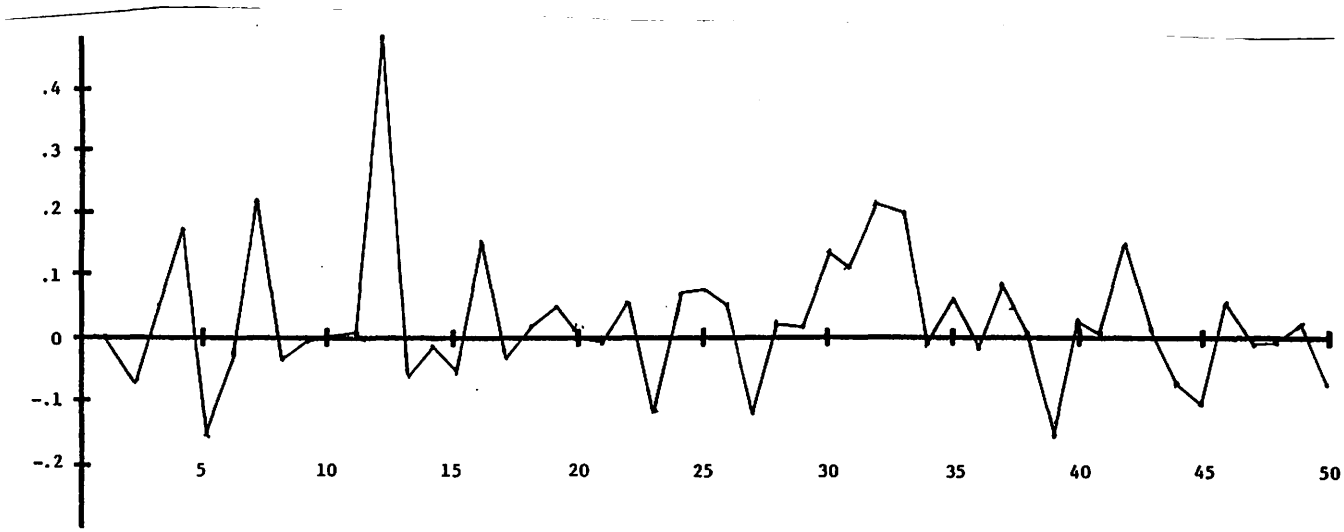


Figure 7.
Index plot for population data full data without UR, corresponding to $\{{}_{12}e_j\}$.

Table 4. Influence statistics for diagnosis data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	$A1O_i^1$	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
1	10.13	12.17	5.59	11.58	1.19	-36.78	-.55	53.72	-1.49	2.49	2.03	-1	-1	.82	Diagnosis (Full)
12	1.86	1.46	240.54	1.32	.42	2.94	.04	28.64	-1.95	3.79	7.30	0	0	.62	
19	1.38	.98	135.22	.84	.32	61.41	.92	61.70	1.97	3.73	7.30	0	0	.46	
16	.17	.21	.12	.23	.17	-23.19	-.35	24.03	-.52	.40	.40	0	0	.40	
17	.12	.13	.10	.13	.14	-7.92	-.12	13.48	.61	.58	.57	0	0	.24	

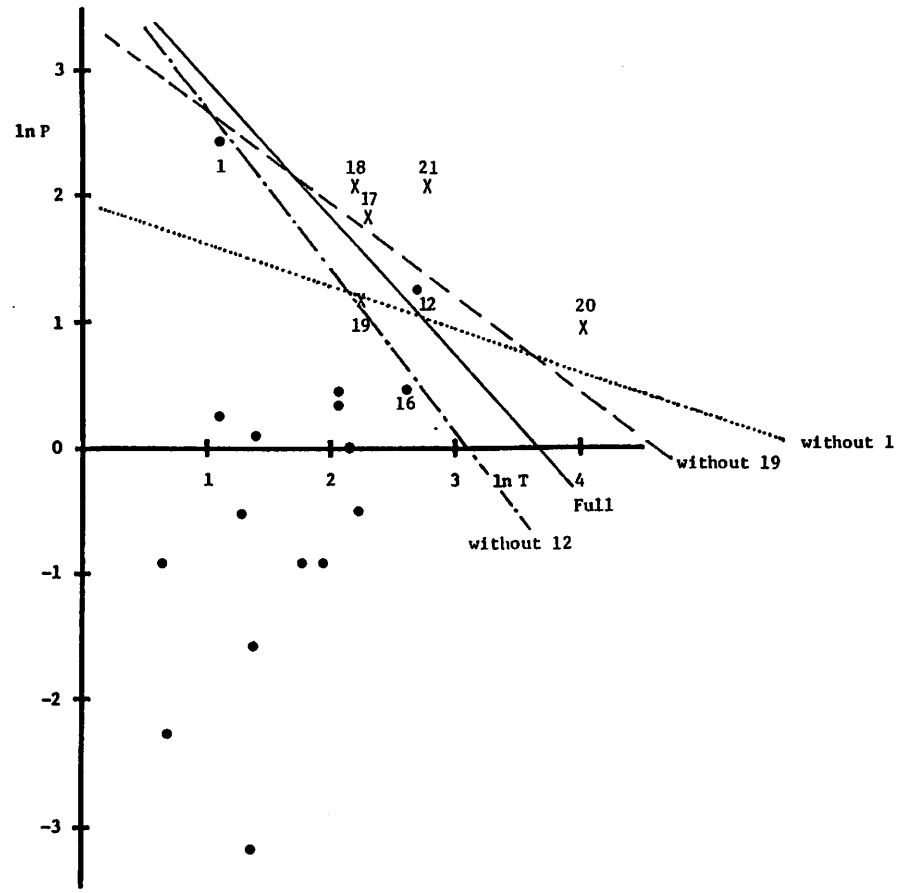


Figure 8.

Diagnosis data. (X) corresponds to a success and (•) corresponds to a failure. Solid line corresponds to the line $x\hat{\beta} = 0$, dashed line to $x\hat{\beta}_{(19)} = 0$, dotted line to $x\hat{\beta}_{(1)} = 0$, and dash-dot-dash to $x\hat{\beta}_{(12)} = 0$.

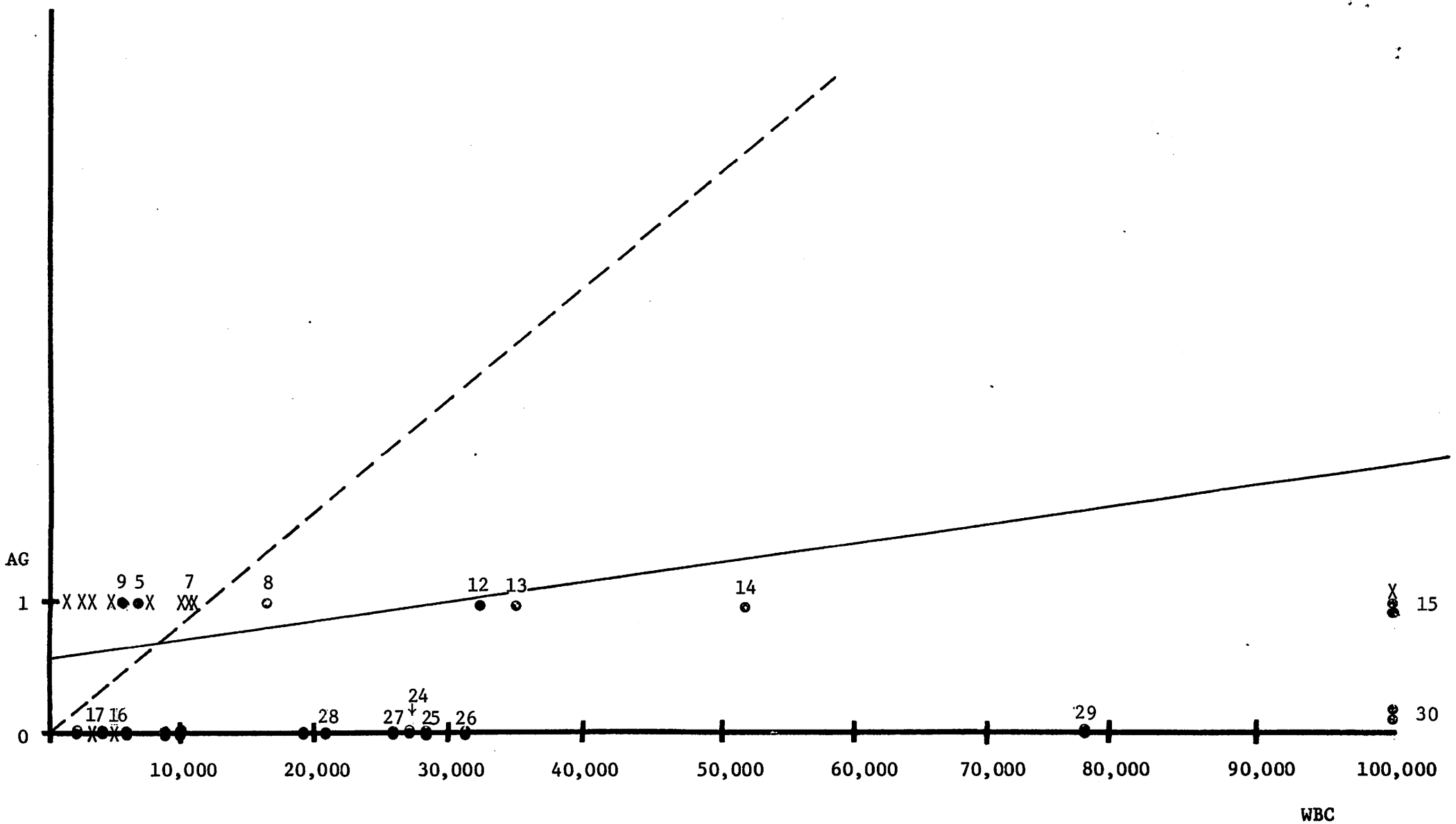


Figure 1

Leukemia Data: (X) indicates "success" and (●) indicates failure. The lines (—) and (---) satisfy $\underset{\sim}{x}\hat{\underset{\sim}{\beta}} = 0$ and $\underset{\sim}{x}\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)} = 0$ where $\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)}$ is the regression coefficients vector determined without case 15 and without a constant. Values below the respective lines are allocated a "failures".

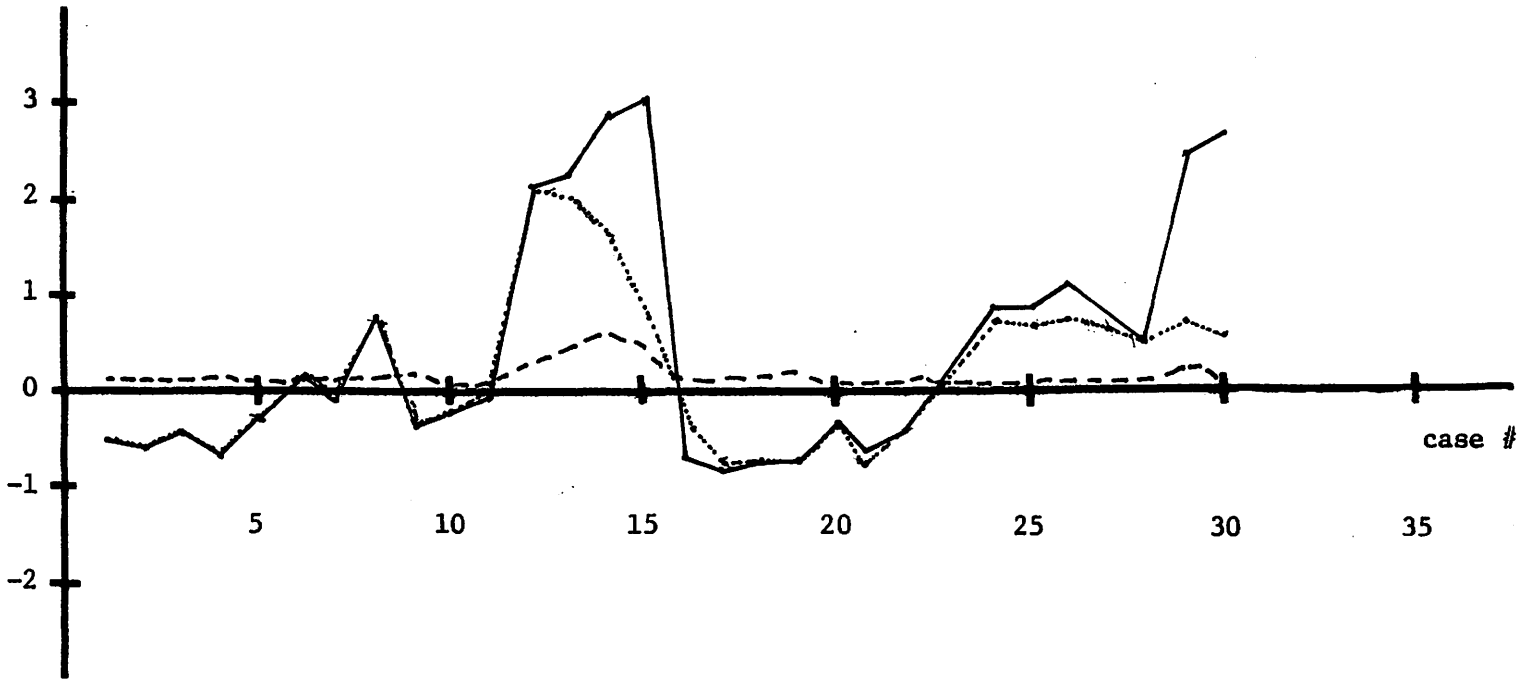


Figure 2.

Index plots for leukemia data corresponding to: $\{_{15}\tilde{\lambda}^1_j\}$ (solid line), $\{_{15}\tilde{e}^1_j\}$ (dotted line) and $\{_{15}g^1_j\}$ (dashed line).

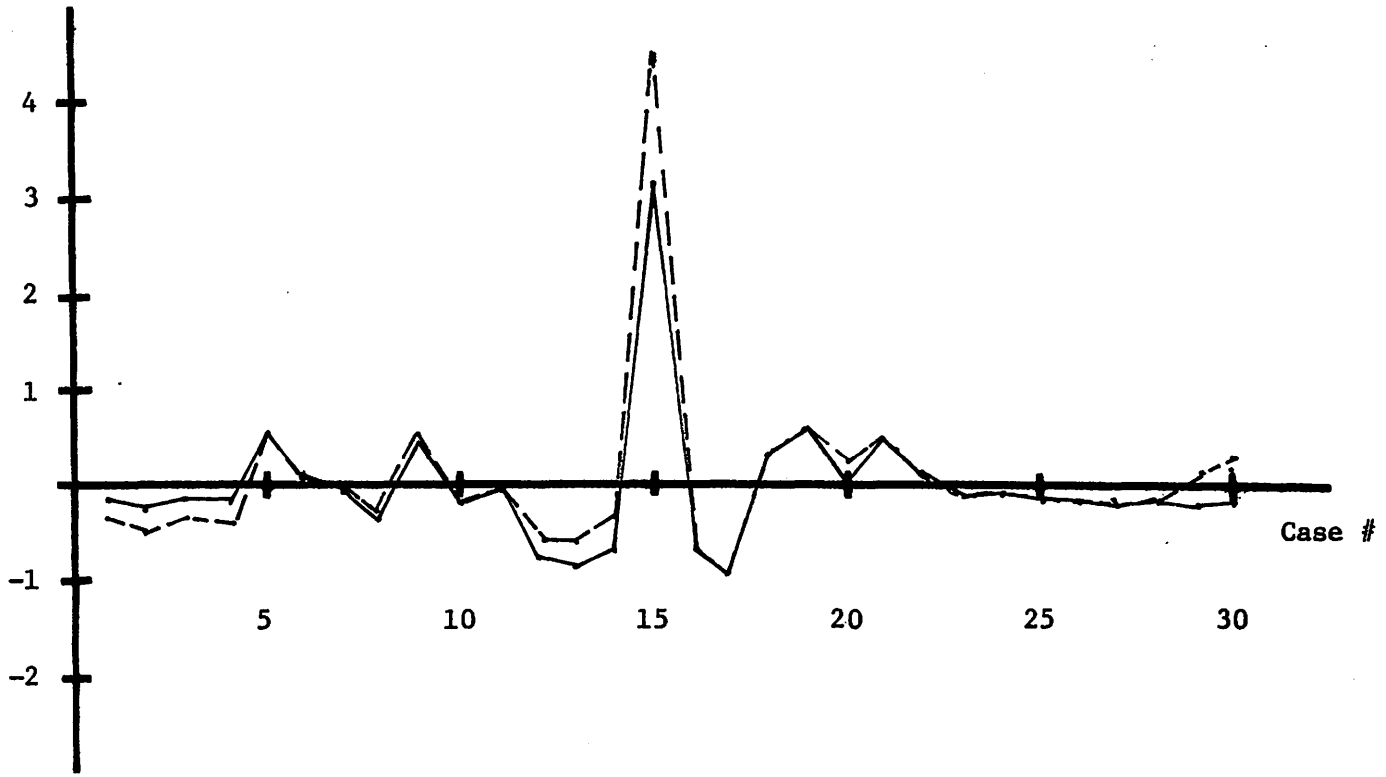


Figure 3.

Index plots for leukemia data corresponding to: $\{\Delta_{15} d_j^1\}$ (solid line) and $\{{}_{15}D_j^1/n_j\}$ (dashed line).

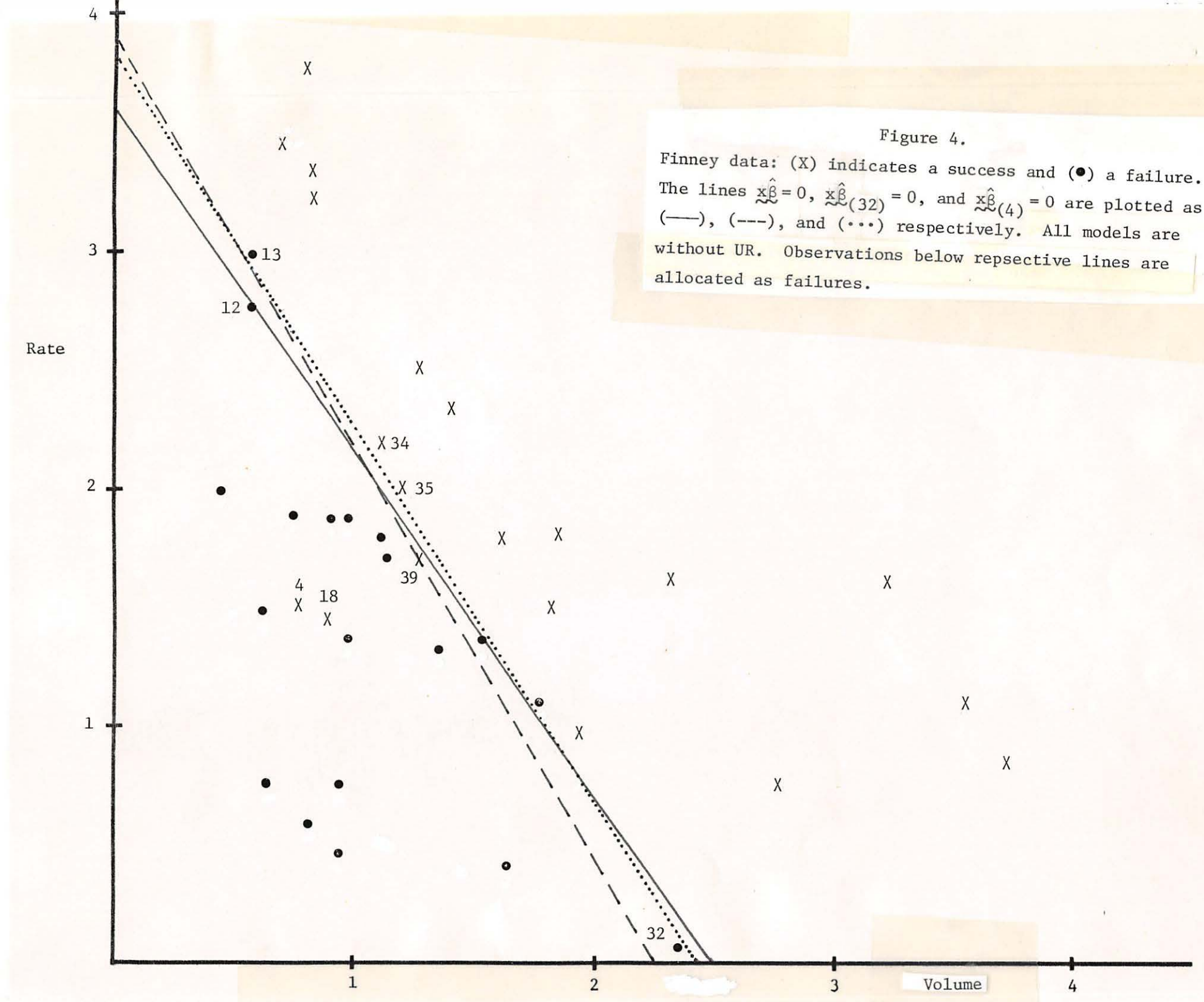


Figure 4.

Finney data: (X) indicates a success and (●) a failure. The lines $x\hat{\beta} = 0$, $x\hat{\beta}_{(32)} = 0$, and $x\hat{\beta}_{(4)} = 0$ are plotted as (——), (---), and (•••) respectively. All models are without UR. Observations below repsective lines are allocated as failures.

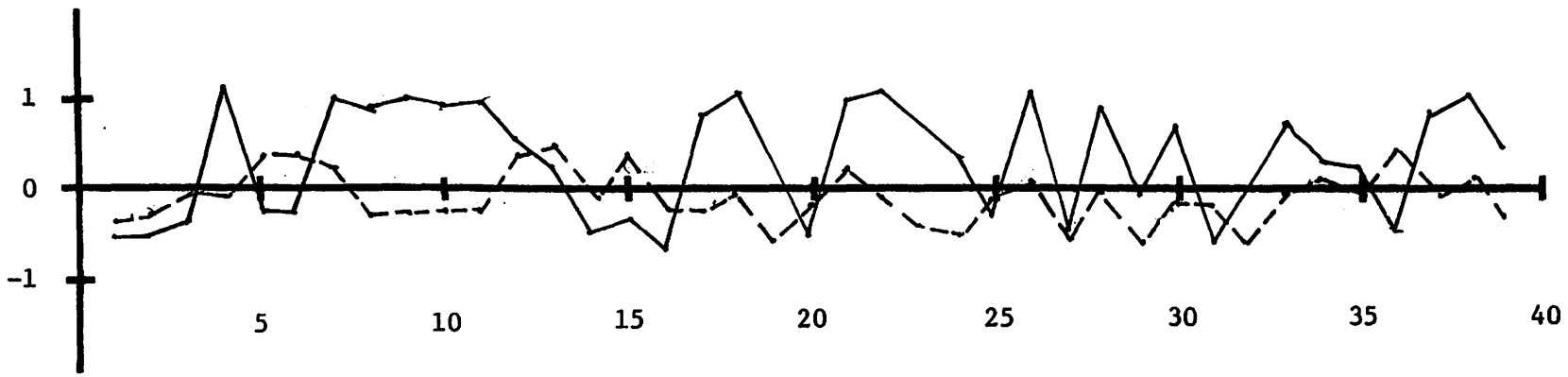


Figure 5.

Index plots for Finney's data corresponding to $\{{}_4\tilde{\lambda}_j\}$ (solid line) and $\{{}_{32}\tilde{\lambda}_j\}$ (dashed line).

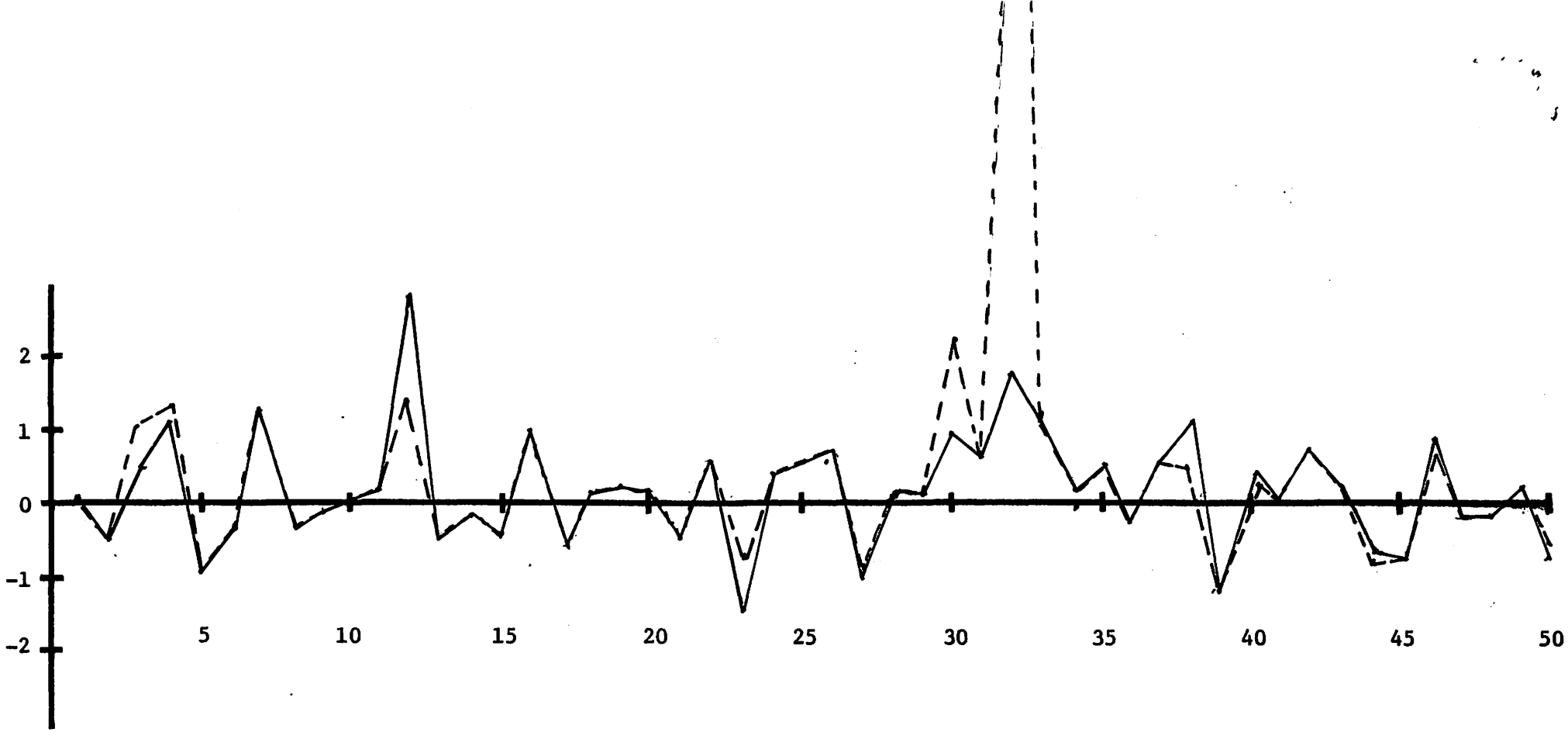


Figure 6.

Index plots for population data (fall data without UR) corresponding to $\{_{12}\tilde{\lambda}_j\}$ (solid line) and $\{_{12}\tilde{e}_j\}$ (dashed line).

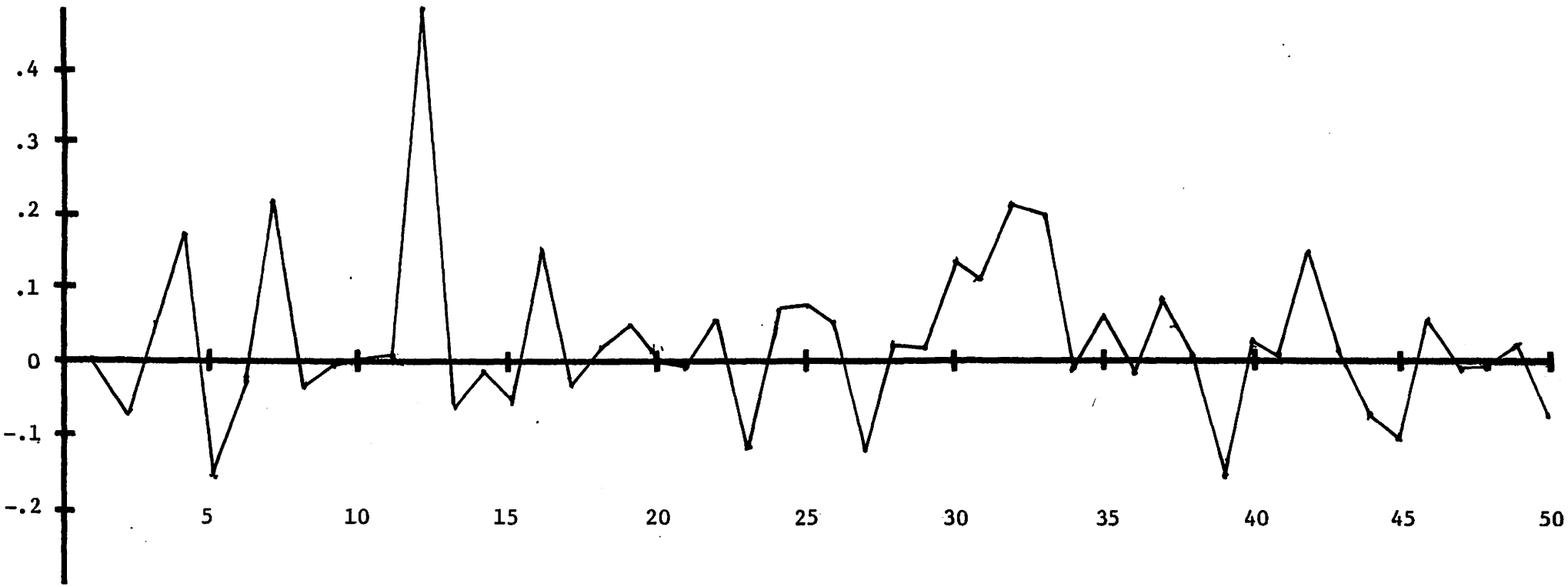


Figure 7.

Index plot for population data full data without UR, corresponding to $\{{}_{12}e_j\}$.

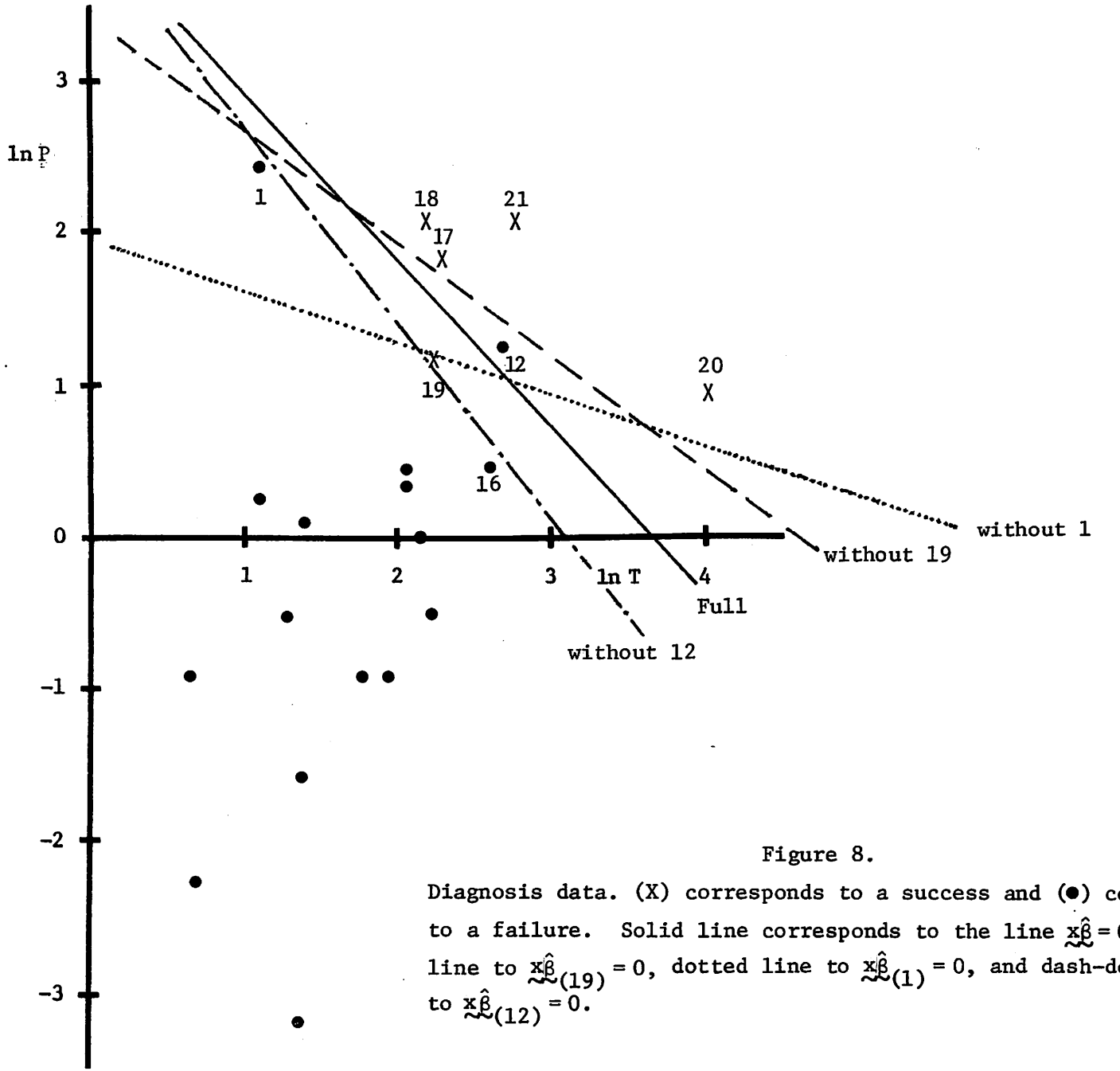


Figure 8.

Diagnosis data. (X) corresponds to a success and (●) corresponds to a failure. Solid line corresponds to the line $\underset{\sim}{x}\hat{\beta} = 0$, dashed line to $\underset{\sim}{x}\hat{\beta}_{(19)} = 0$, dotted line to $\underset{\sim}{x}\hat{\beta}_{(1)} = 0$, and dash-dot-dash to $\underset{\sim}{x}\hat{\beta}_{(12)} = 0$.

Table 1. Influence measures for leukemia data

Case #	D_i^1	LD_i^1	LD_i	DIV_i^1	ED_i^1	LO_i^1	SLO_i^1	ALO_i^1	$\Delta_i\chi^*$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCI_i^1$	$\hat{v}_{ii}$	Data set
15	9.94	4.85	10.72	2.67	.59	16.27	1.13	28.07	2.32	4.74	8.14	0	0	.65	Leukemia (Full)
17	.51	.45	.71	.41	.22	7.50	.52	8.42	2.13	3.70	3.77	0	0	.10	
16	.50	.45	.68	.41	.22	7.58	.53	8.40	2.17	3.77	3.84	0	0	.10	
9	.21	.21	.24	.20	.20	-2.77	-.19	4.75	-1.54	2.51	2.52	-2	0	.08	
5	.21	.21	.22	.19	.20	-2.77	-.19	4.65	-1.53	2.48	2.49	-2	0	.08	
9	.52	.47	.67	.44	.25	5.79	.12	17.57	-2.17	3.79	3.85	0	0	.11	Leukemia (without case 15, without constant)
5	.47	.43	.57	.41	.25	3.09	.06	14.42	-2.04	3.54	3.59	0	0	.11	
8	.16	.14	.13	.18	.16	-16.78	-.35	16.91	-.67	.69	.68	0	0	.34	
16	.12	.11	.16	.10	.10	16.17	.21	18.33	1.66	2.69	2.71	0	0	.19	
7	.12	.12	.12	.12	.15	7.22	.15	7.22	.87	1.10	1.10	0	0	.15	

* $\Delta_i\chi = \chi_i/(1-\hat{v}_{ii})^{\frac{1}{2}}$

Tableau n° [illegible] Résultats des mesures aérodynamiques

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 2. Influence statistics for Finney's data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	ALO_i^1	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
4	1.05	.86	1.59	.79	.30	4.95	.27	28.90	3.88	6.40	6.70	0	0	.07	Finney (Full)
18	.91	.76	1.47	.69	.28	5.29	.29	26.58	3.58	5.96	6.19	0	0	.07	
32	.55	.55	.58	.54	.33	-7.06	-.39	12.25	-1.04	1.45	1.45	-3	0	.33	
13	.49	.45	.59	.44	.27	-5.29	-.29	10.04	-1.59	2.68	2.71	-1	0	.16	
12	.20	.20	.22	.20	.19	-3.30	-.18	8.31	-1.03	1.45	1.64	-1	0	.16	
13	3.59	2.70	8.17	2.34	.55	-39.84	-1.21	100.74	-2.00	4.15	4.97	0	0	.47	Finney (without cases 4 and 18)
32	1.68	1.69	1.97	1.62	.58	-15.39	-.47	28.45	-1.14	1.64	1.68	-1	0	.56	
39	1.30	.90	*	.77	.27	8.03	.24	163.82	2.66	4.98	*	0	0	.16	
35	.18	.19	.18	.20	.20	13.38	.41	34.05	.87	1.10	1.09	0	0	.19	
34	.16	.17	.16	.18	.19	12.16	.37	40.96	.73	.81	.10	0	0	.23	

* indicates that estimates would not converge.

Table 3. Influence statistics for population data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	AlO_i^1	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
12	7.56	6.41	5.61	5.54	.84	17.63	.93	33.14	1.53	2.68	2.63	1	-3	.76	Population (Full)
35	1.17	1.03	1.44	.97	.39	8.56	.45	23.61	2.57	4.77	4.90	1	-1	.15	
47	1.25	1.11	1.45	1.05	.42	2.67	-.14	21.71	-2.11	3.98	4.08	-1	0	.22	
37	1.10	.97	1.41	.91	.37	-9.16	-.48	21.38	-2.83	5.13	5.27	0	0	.12	
9	.43	.40	.51	.38	.23	6.86	.36	16.02	2.62	4.43	4.47	0	-2	.06	
12	8.06	6.27	5.46	5.48	.83	17.10	.91	34.29	1.73	3.38	3.29	1	-2	.73	Population (without UR)
47	1.26	1.13	1.46	1.07	.43	3.36	-.18	21.58	-2.07	3.91	3.99	-1	1	.23	
35	.43	.40	.45	.39	.26	6.74	.36	13.13	1.84	3.16	3.18	1	0	.11	
9	.42	.39	.48	.37	.22	6.76	.36	15.76	2.71	4.55	4.59	0	0	.05	
37	.89	.81	1.07	.77	.34	16.86	-.44	16.86	-2.53	4.56	4.64	0	0	.12	
47	1.03	.90	1.33	.84	.36	-3.43	-.18	18.68	-2.11	3.90	4.03	-1	3	.11	Population (without constant, UR, or case 12)
37	.73	.66	.90	.63	.30	-6.77	-.35	17.30	-2.31	4.12	4.20	0	0	.12	
35	.57	.52	.66	.50	.28	9.15	.47	15.96	2.22	3.89	3.94	1	0	.11	
9	.45	.41	.54	.40	.23	8.40	.44	17.68	2.85	4.76	4.80	1	0	.05	
24	.34	.33	.36	.32	.22	-6.03	-.31	10.06	-1.61	2.68	2.70	0	0	.09	

Table 4. Influence statistics for diagnosis data.

Case #	D_i^1	LD_i^1	LD_i	DIV^1	ED_i^1	LO_i^1	SLO_i^1	ALO_i^1	$\Delta_i X$	$\Delta_i D^1$	$\Delta_i D$	$NCCO_i^1$	$NCCl_i^1$	$\hat{v}_{ii}$	Data set
1	10.13	12.17	5.59	11.58	1.19	-36.78	-.55	53.72	-1.49	2.49	2.03	-1	-1	.82	Diagnosis (Full)
12	1.86	1.46	240.54	1.32	.42	2.94	.04	28.64	-1.95	3.79	7.30	0	0	.62	
19	1.38	.98	135.22	.84	.32	61.41	.92	61.70	1.97	3.73	7.30	0	0	.46	
16	.17	.21	.12	.23	.17	-23.19	-.35	24.03	-.52	.40	.40	0	0	.40	
17	.12	.13	.10	.13	.14	-7.92	-.12	13.48	.61	.58	.57	0	0	.24	

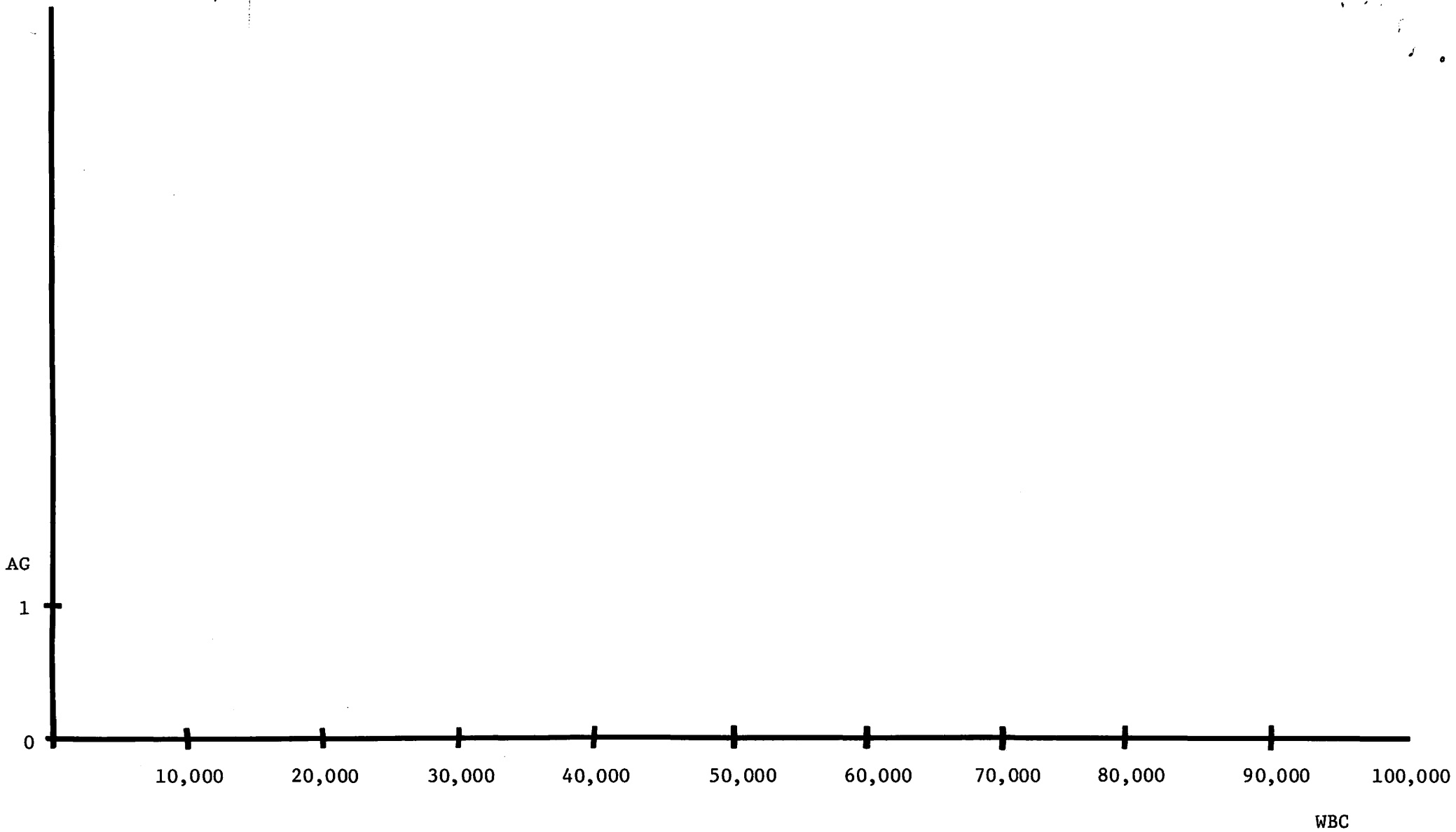


Figure 1

Leukemia Data: (X) indicates "success" and (●) indicates failure. The lines (—) and (---) satisfy $\underset{\sim}{x}\hat{\underset{\sim}{\beta}} = 0$ and $\underset{\sim}{x}\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)} = 0$ where $\hat{\underset{\sim}{\beta}}^{\bar{c}}_{(15)}$ is the regression coefficients vector determined without case 15 and without a constant. Values below the respective lines are allocated a "failures".

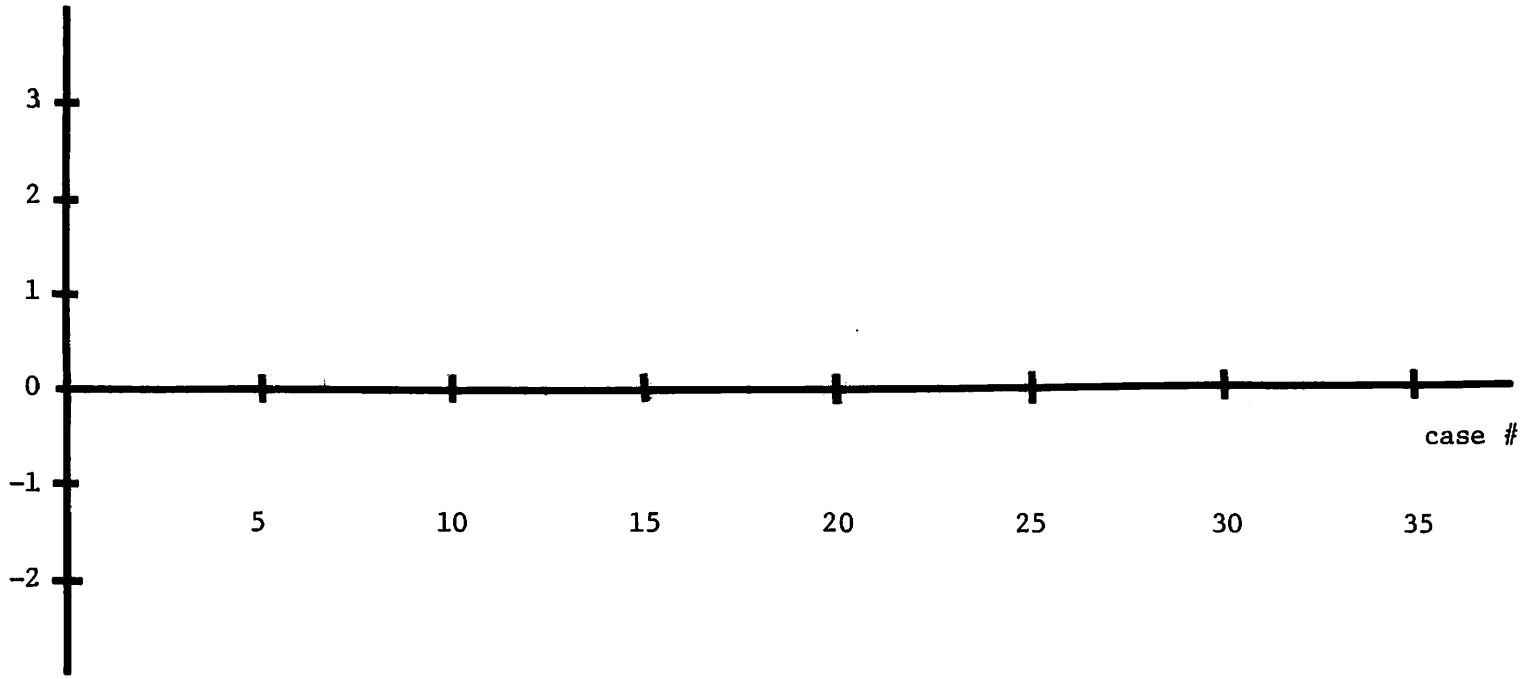


Figure 2.

Index plots for leukemia data corresponding to: $\{15\tilde{\lambda}_j^1\}$ (solid line), $\{15\,e_j^1\}$ (dotted line) and $\{15\hat{y}_i^1\}$ (dashed line).

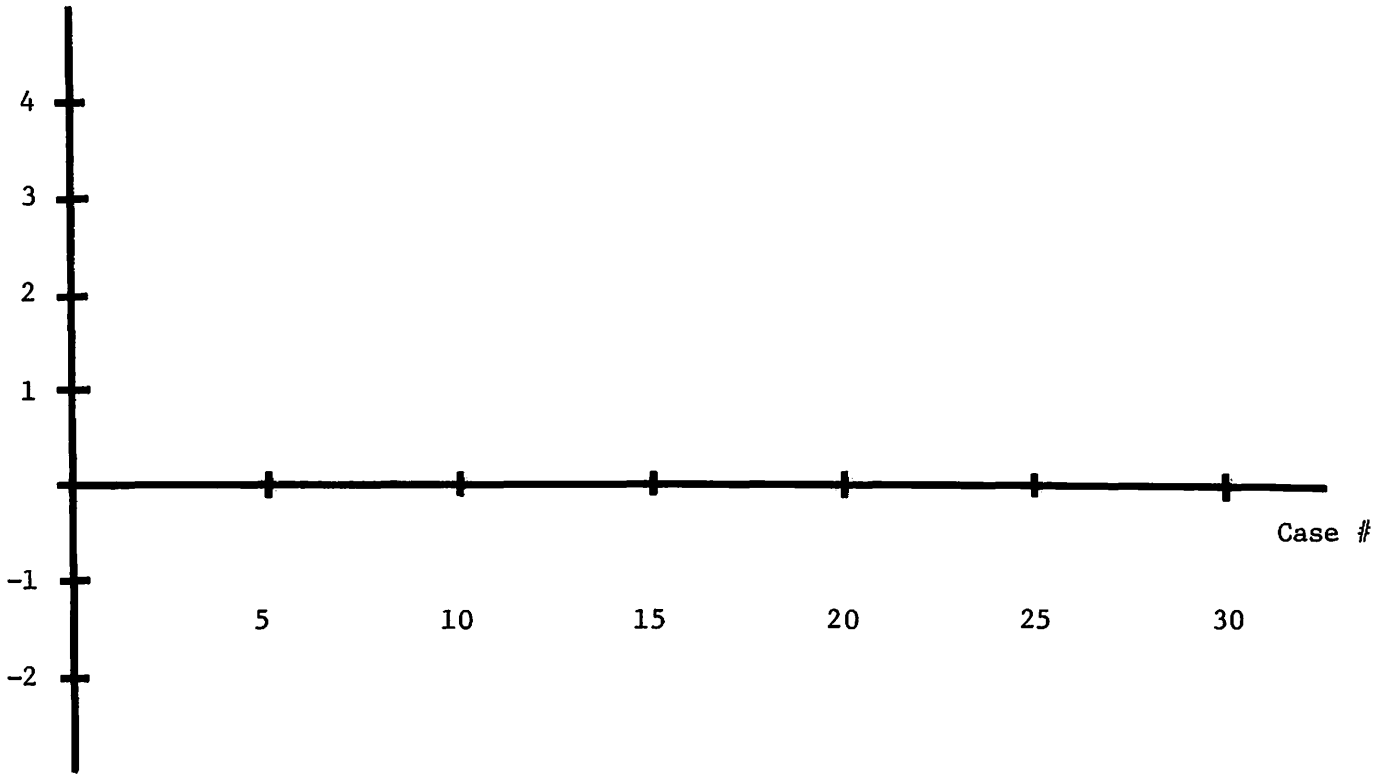


Figure 3.

Index plots for leukemia data corresponding to: $\{\Delta_{15} d_j^1\}$ (solid line) and $\{{}_{15}\Delta_j^1/n_j\}$ (~~dotted line~~).

↓

dashed

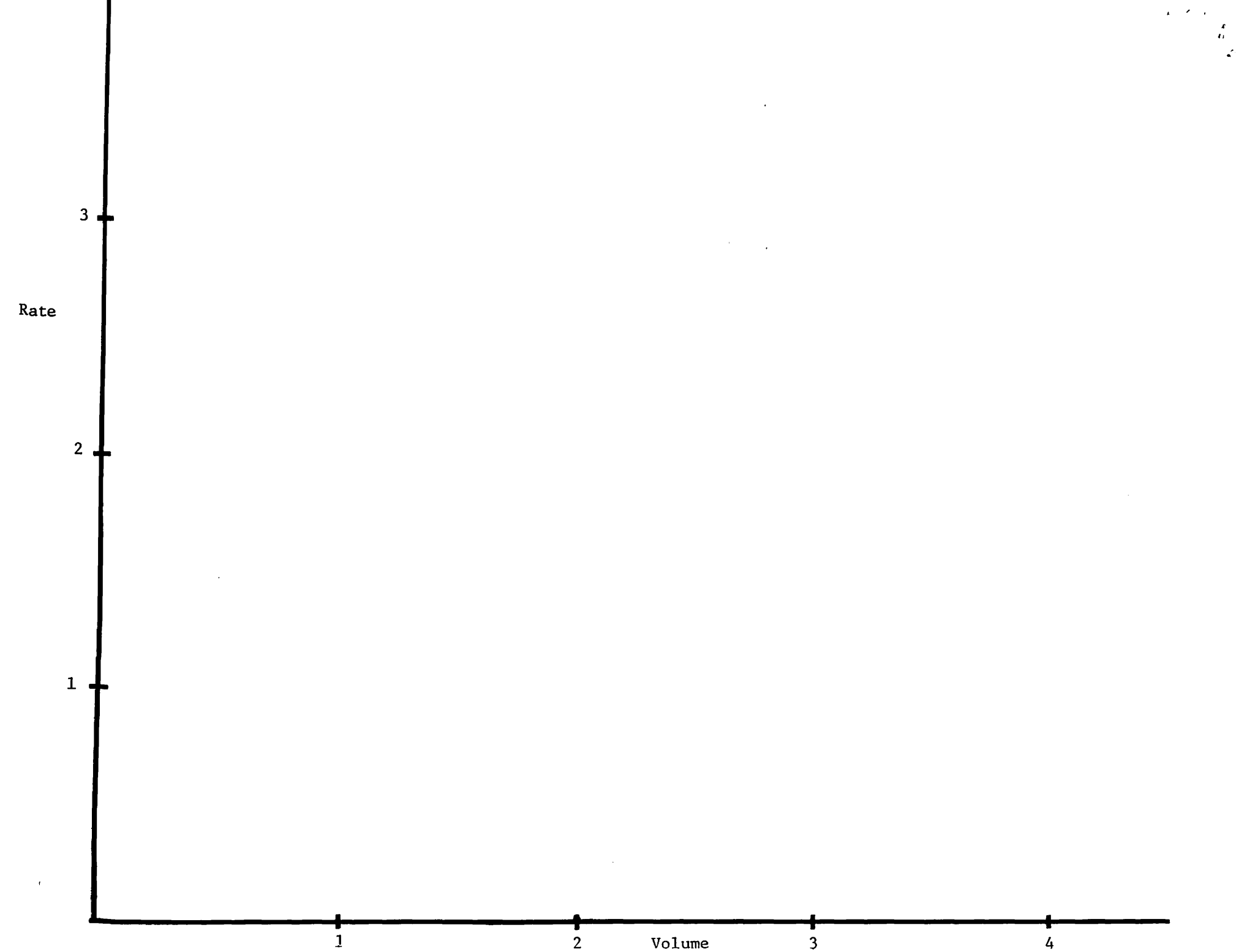

3
Rate
2
1
1
2
Volume
3
4

Figure 5.

Index plots for Finney's data corresponding to $\{{}_4\tilde{\lambda}_j\}$ (solid line) and $\{{}_{32}\tilde{\lambda}_j\}$ (dashed line).